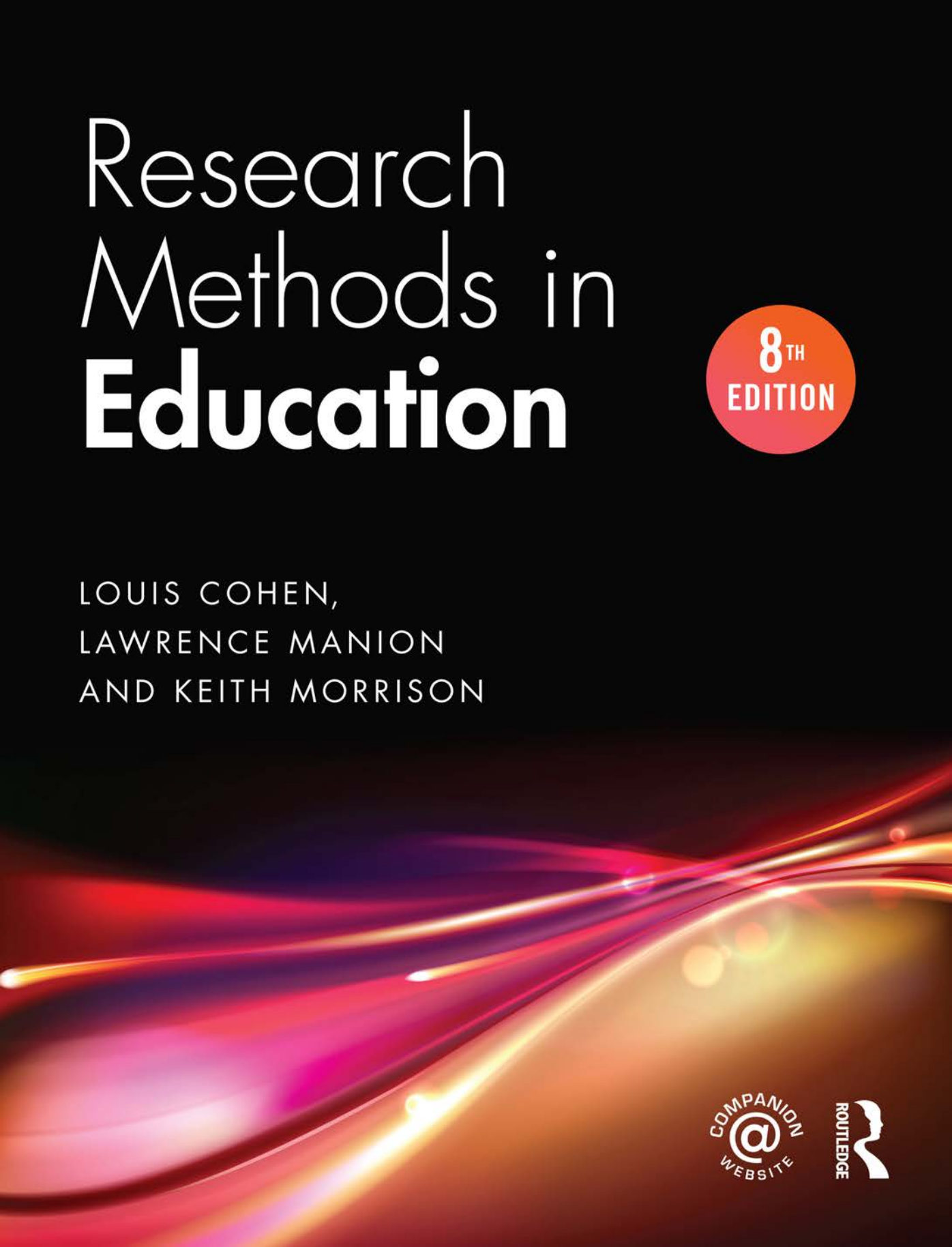
Research Methods in **Education**

8TH EDITION

LOUIS COHEN,
LAWRENCE MANION
AND KEITH MORRISON



Very much still the key text for 'all' education students and researchers. Cohen *et al*. continue to update *Research Methods in Education*, with new theoretical, ethical, virtual and mixed methods information. It's worth noting the impressive web page and links to materials for all chapters which is still the benchmark when looking at the competition for books in this area of social and education research.

Dr Richard Race, Senior Lecturer in Education, Roehampton University, UK

A clear enhancement on the already well-established text. The new edition addresses an important need to explain research design and question setting in more detail, helping guide the newcomer through the research process from inception through analysis to reporting.

David Lundie, Associate Professor of Education, University of St Mark & St John, UK

Research Methods in Education is a unique book for everybody who has to undertake educational research projects. The book gives an in depth understanding of quantitative and qualitative research designs and offers a practical guide for data collection and data analysis. It is an essential 'friend' for teachers and students from various disciplines who are not familiar with social science research.

Dr Ellen P. W. A. Jansen, Associate Professor, Teacher Education, University of Groningen, The Netherlands

Research Methods in Education continues to offer an excellent route map, a well-structured and inspiring travel guide, for students engaging in research. It works across levels, and while it provides clarity for the beginning researcher there is plenty here to aid the seasoned researcher with an open mind to new approaches and emerging practices. A superb text that provides guidance for my own research as well as for students and partners in research projects.

Peter Shukie, Lecturer in Education Studies and Academic Lead in Digital Innovation, University Centre at Blackburn College, UK

Research Methods in Education is, besides being my personal favorite research methods book, a deep as well as a broad handbook useful both for undergraduate teacher education students as well as researchers and PhD students within educational sciences. In this new edition, new chapters are added emphasising both quantitative and qualitative methods in combination with thought-through discussions about how to mix them. The book can be used when planning a project and then throughout the whole research process and is therefore a complete methods book.

Karolina Broman, Senior Lecturer in Chemistry Education, Umeå University, Sweden

Comprehensive, well written and relevant: the eighth edition of *Research Methods in Education* offers the background for methods courses at different levels. The new edition keeps the strong focus on education studies. Excellent extensions will make the book an even more popular basis for classes on both qualitative and quantitative methods.

Felix Weiss, Assistant Professor for Sociology of Education, Aarhus University, Denmark

Research Methods in Education, Eighth Edition is an up-to-date, one-stop shop, taking education research students from conceptualization to presentation. With this book on your library shelf, you are good to go.

Dr Fiona McGarry, Lecturer in Research Methods, University of Dundee, UK

The eighth edition of *Research Methods in Education* contains a wealth of up-to-the-minute information and guidance on educational research which will be of immense value to researchers at all stages of their careers and across the education domain from early years settings to higher education. As research and education move into increasingly fluid and complex dimensions, *Research Methods in Education* will support students, researchers and practitioners in charting a course through these changing waters as they seek to create new knowledge about effective teaching and deepen our understanding of how learners learn.

Julia Flutter, A Director of the Cambridge Primary Review Trust, Faculty of Education, University of Cambridge, UK

As a doctoral supervisor I know that my students routinely return to *Research Methods in Education* as they develop their own research projects. This text has always been a mainstay on our reading lists but this new edition now features additional research topics and new perspectives on a wider range of research methods. As with previous editions this book is clearly organised and well written and appeals to a wide audience of experienced and novice researchers alike.

Dr Val Poultney, Associate Professor, University of Derby, UK

Research Methods in Education

This thoroughly updated and extended eighth edition of the long-running bestseller *Research Methods in Education* covers the whole range of methods employed by educational research at all stages. Its five main parts cover: the context of educational research; research design; methodologies for educational research; methods of data collection; and data analysis and reporting. It continues to be the go-to text for students, academics and researchers who are undertaking, understanding and using educational research, and has been translated into several languages. It offers plentiful and rich practical advice, underpinned by clear theoretical foundations, research evidence and up-to-date references, and it raises key issues and questions for researchers planning, conducting, reporting and evaluating research.

This edition contains new chapters on:

- Mixed methods research
- The role of theory in educational research
- Ethics in Internet research
- Research questions and hypotheses
- Internet surveys
- Virtual worlds, social network software and netography in educational research
- Using secondary data in educational research
- Statistical significance, effect size and statistical power
- Beyond mixed methods: using Qualitative Comparative Analysis (QCA) to integrate cross-case and within-case analyses.

Research Methods in Education is essential reading for both the professional researcher and anyone involved in educational and social research. The book is supported by a wealth of online materials, including PowerPoint slides, useful weblinks, practice data sets, downloadable tables and figures from the book, and a virtual, interactive, self-paced training programme in research methods. These resources can be found at: www.routledge.com/cw/cohen.

Louis Cohen is Emeritus Professor of Education at Loughborough University, UK.

Lawrence Manion was Principal Lecturer in Music at Manchester Metropolitan University, UK.

Keith Morrison is Professor and Advisor for Institutional Development at Macau University of Science and Technology, China.

Research Methods in Education

Eighth edition

Louis Cohen, Lawrence Manion and
Keith Morrison

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Eighth edition published 2018
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

and by Routledge
711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

Seventh edition published by Routledge 2011.

Library of Congress Cataloging-in-Publication Data
Names: Cohen, Louis, 1928– author. | Manion, Lawrence, author. | Morrison, Keith (Keith R. B.) author.
Title: Research methods in education / Louis Cohen, Lawrence Manion and Keith Morrison.
Description: Eighth edition. | New York: Routledge, 2018. | Includes bibliographical references and index.
Identifiers: LCCN 2017015256| ISBN 9781138209862 (hardback) | ISBN 9781138209886 (paperback) | ISBN 9781315456539 (ebook)
Subjects: LCSH: Education–Research–Great Britain.
Classification: LCC LB1028.C572 2018 | DDC 370.72–dc23
LC record available at https://lccn.loc.gov/2017015256

ISBN: 978-1-138-20986-2 (hbk)
ISBN: 978-1-138-20988-6 (pbk)
ISBN: 978-1-315-45653-9 (ebk)

Typeset in Times New Roman
by Wearset Ltd, Boldon, Tyne and Wear

Visit the companion website: www.routledge.com/cw/cohen

Contents

Figures

Tables

Boxes

Contributors

Richard Bell, PhD, Honorary staff member and formerly Associate Professor in the Department of Psychological Sciences, University of Melbourne, has written Chapter 29: 'Personal constructs'.

Barry Cooper, PhD, Emeritus Professor of Education in the School of Education, University of Durham, has jointly written Chapter 45: 'Beyond mixed methods: using Qualitative Comparative Analysis (QCA) to integrate cross-case and within-case analyses'.

Judith Glaesser, PhD, Research Associate for Evaluation in the School of Education at Eberhard Karls Universität Tübingen, has jointly written Chapter 45: 'Beyond mixed methods: using Qualitative Comparative Analysis (QCA) to integrate cross-case and within-case analyses'.

Jane Martin, PhD, Professor of Social History of Education and Head of the Department of Education and Social Justice, University of Birmingham, has written Chapter 16: 'Historical and documentary research', and is currently conducting research on Caroline Benn.

Stewart Martin, PhD, Professor of Education at the School of Education and Social Sciences, University of Hull, has written Chapter 23: 'Virtual worlds, social network software and netography in educational research'.

Carmel O'Sullivan, PhD, Associate Professor of Education and Head of School of Education at Trinity College Dublin, has written Chapter 30: 'Role-play and research'.

Harsh Suri, PhD, Senior Lecturer in Learning Futures in the Faculty of Business and Law at Deakin University, has written Chapter 21: 'Meta-analysis, systematic reviews and research syntheses'.

Preface to the eighth edition

We are indebted to Routledge for the opportunity to produce an eighth edition of our book *Research Methods in Education*. The book continues to be received very favourably worldwide; it is the standard text for many courses in research methods and has been translated into several languages.

The eighth edition contains much new material, including entirely new chapters on:

- Paradigms in educational research
- Mixed methods research
- The role of theory in educational research
- Ethics in Internet research
- Research questions and hypotheses
- Historical and documentary research
- Internet surveys
- Meta-analysis, research syntheses and systematic reviews
- Virtual worlds, social network software and netography in educational research
- Using secondary data in educational research
- Statistical significance, effect size and statistical power
- Beyond mixed methods: using Qualitative Comparative Analysis (QCA) to integrate cross-case and within-case analyses.

Whilst retaining the best features of the former edition, the reshaping, updating and new additions undertaken for this new volume now mean that the book covers a greater spread of issues than the previous editions, and in greater depth, catching the contemporary issues and debates in the field. In particular, the following new material has been included:

Part 1:
- Post-positivism, post-structuralism and postmodernism
- Constructionism in educational research
- Subjectivity and objectivity in educational research
- Mixed methods research
- Paradigms, ontology and epistemology in mixed methods research
- Working with mixed methods research
- Stages in mixed methods research
- Value-neutrality in educational research
- The role of theory in educational research
- Types and meanings of theory
- Worked examples of causation in educational research

Part 2:
- Regulatory contexts of ethics
- Sponsored research
- Ethical codes and their limitations
- Ethics and the quality of research
- Power and position

- Reciprocity
- Ethics in data analysis, reporting and dissemination
- Key ethical issues in Internet research
- Challenges to privacy and informed consent in Internet research
- Public and private matters in Internet research
- Ethical codes in Internet research
- Choosing a research project
- Deriving and devising research questions
- Different kinds of research question
- Organizing research questions
- The need for warrants in educational research
- Statistical power in sampling issues
- Sampling in mixed methods research
- Effects of sensitive research on the researcher

Part 3:

- Autoethnography
- Virtual ethnography
- Reflexivity
- Historical and documentary research
- Survey questions
- Low response, non-response and missing data in surveys
- Constructing Internet-based surveys
- Ethical issues in Internet-based surveys
- Typology of case studies
- Generalization in case study
- What makes a good case study researcher?
- Randomized controlled trials
- The importance of randomization
- Concerns about randomized controlled trials
- The limits of averages in randomized controlled trials
- Null hypothesis significance testing
- Participatory action research
- Ethical issues in action research

Part 4:

- Considering the demands on the respondent in questionnaire construction
- Administering questionnaires
- Planning and conducting interviews
- Prompts and probes in interviews
- Interviewing children
- Group interviewing
- Telephone interviewing
- Online interviewing
- Key issues in observations
- Video observations
- Using secondary data in educational research
- Sources and types of secondary data
- Advantages of, and challenges in, using secondary data
- Ethical issues in using secondary data
- Examples of secondary data analysis
- Working with secondary data
- Photo-elicitation

- Provision of images in educational research
- Video and moving images in educational research
- Ethical practices in visual research

Part 5:

- Elements of qualitative data analysis
- Making sense of qualitative data
- Computer Assisted Qualitative Data Analysis (CAQDAS)
- Examples of CAQDAS
- Reflexivity in CAQDAS
- Strengths and weaknesses of CAQDAS
- Advances in CAQDAS
- Ways of organizing and presenting qualitative data analysis
- Examples of coding qualitative data with software (CAQDAS)
- Concerns about coding
- Content analysis with software (CAQDAS)
- Worked examples of using software in analysing visual data (CAQDAS)
- Challenges in interpreting visual images
- Analysing moving images
- Versions of, stages in and concerns about grounded theory
- Moderator and mediator variables
- Confidence intervals
- Concerns about statistical significance
- Hypothesis testing and null hypothesis significance testing
- Statistical power
- Coping with missing data
- 'Safety checks' and assumptions when using statistics (for all the statistics addressed)
- Command sequences for running statistics in the Statistical Package for the Social Sciences (SPSS)
- Reporting statistical analysis
- Cluster analysis
- What to look for in factor analysis output
- Additions to guidance charts when choosing statistics
- Using Qualitative Comparative Analysis (QCA) to integrate cross-case and within-case analyses
- Starting from quantitative and qualitative stances in QCA
- Ragin's QCA
- Worked examples of QCA

A signal feature of this edition is the inclusion of very many extensively worked examples and more figures, diagrams and graphics to illustrate and summarize key points clearly. Several of the tables in Part 5 include SPSS and NVivo output, so that readers can check their own SPSS and NVivo analysis against the examples provided.

To accompany this volume, a companion website provides a comprehensive range of materials to cover all aspects of research (including summaries of every chapter on PowerPoint slides), exercises and examples, explanatory material and further notes, website references, SPSS data files, QSR NVivo data files, together with further statistics and statistical tables. These are indicated in the book.

This book stands out for its practical advice that is securely rooted in theory and up-to-date discussion from a range of sources. We hope that it will continue to constitute the first 'port of call' for educational researchers and continue to be the definitive text in its field.

Acknowledgements

Our thanks are due to the following publishers and authors for permission to include materials in the text:

American Educational Research Association, for words from Strike, K. A., Anderson, M. A., Curren, R., van Geel, T., Pritchard, I. and Robertson, E. (2000) *Ethical Standards of the American Educational Research Association 2000*. Washington, DC: American Educational Research.

American Psychological Association, for words from American Psychological Association (2010) *Publication Manual of the American Psychological Association* (sixth edition). Washington, DC: Author.

Association of Internet Researchers, for words from Association of Internet Researchers (2012) *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0)*.

Beamish Museum, UK, for photograph No. 29474.

Bloomsbury Publishing, Plc, for words from Hammersley, M. © (2013) *What Is Qualitative Research?* Bloomsbury Academic, an imprint of Bloomsbury Publishing Plc; Kettley, N. © (2012) *Theory Building in Educational Research*. Continuum, used by permission of Bloomsbury Publishing Plc; Wellington, J. © (2015) *Doing Qualitative Educational Research: A Personal Guide to the Research Process* (second edition). Continuum, used by permission of Bloomsbury Publishing Plc. For anonymous, third-party interview words reported in Walford, G. © (2001) *Doing Qualitative Educational Research: A Personal Guide to the Research Process*. Continuum, used by permission of Bloomsbury Publishing Plc.

British Educational Research Association, for words from British Educational Research Association (2011) *Ethical Guidelines for Educational Research*. London: British Educational Research Association.

British Medical Journal Publishing Group Ltd, for material from Curr, D. (1994) Role play. *British Medical Journal*, 308 (6930), p. 725.

British Psychological Society, for words from British Psychological Society (2013) *Ethics Guidelines for Internet-Mediated Research*. Leicester, UK: British Psychological Society; British Psychological Society (2014) *Code of Human Research Ethics*. Leicester, UK: British Psychological Society.

British Sociological Association, for words from British Sociological Association (2002) *Statement of Ethical Practice*. Durham, UK: British Sociological Association. Reproduced with permission from © The British Sociological Association.

Brookshire, R. G. and Bartlett, J. E., for material from Bartlett, J. E., II, Kotrlik, J. W. and Higgins, C. C. (2001) Organizational research: determining appropriate sample size in survey research. *Information Technology, Learning and Performance Journal*, 19 (1), pp. 43–50.

Cambridge University Press, for words from Strauss, A. L. (1987) *Qualitative Analysis for Social Scientists*. Cambridge: Cambridge University Press.

Economic and Social Research Council, for words from Economic and Social Research Council (2015) *ESRC Framework for Research Ethics*. Swindon, UK: Economic and Social Research Council.

HarperCollins Publishers Ltd, for materials from Cohen, L. and Holliday, M. (1979) *Statistics for Education and Physical Education*.

Harvard Education Publishing Group, for words from Carver, R. P. (1978) The case against statistical significance testing. *Harvard Educational Review*, 48 (3), pp. 378–99.

Higher Education Research and Development and B. Grant (Editor), for words from Hammersley, M. (2012) Troubling theory in case study research. *Higher Education Research and Development*, 31 (3), pp. 393–405.

Hindawi Publishing Corporation, for words from Leshem, S. (2012) The group interview experience as a tool for admission to teacher education. *Education Research International*, Article ID 876764. Available from: http://dx.doi.org/10.1155/2012/876764.

Human Kinetics Inc., for words from Sparkes, A. C. (2000) Autoethnography and narratives of self: reflections on criteria in action. *Sociology of Sport Journal*, 17 (1), pp. 21–43.

Jean McNiff and September Books for words from McNiff, J. (2010) *Action Research for Professional Development: Concise Advice for New and Experienced Action Researchers*. Poole, UK: September Books.

John Wiley & Sons, for words from Dyer, C. (1995) *Beginning Research in Psychology*. Oxford: Blackwell.

Labaree, R. V. and *University of Southern California*, for words from Labaree, R. V. (2013) *Organizing Your Social Sciences Research Paper: Types of Research Designs*. USC Libraries Research Guides. Los Angeles, CA: University of Southern California.

McGraw-Hill, for words from Denscombe, M. (2014) *The Good Research Guide* (fourth edition). Maidenhead, UK: Open University Press.

Mosaic Books and *PRIA: Society for Participatory Research in Asia*, for words from Tandon, R. (ed.) (2005) *Participatory Research: Revisiting the Roots*.

Palgrave Macmillan, for words from Torgerson, C. J. and Torgerson, D. J. (2008) *Designing Randomised Trials in Health, Education and the Social Sciences*. Houndmills, UK: Palgrave Macmillan.

Penguin Random House, for excerpts from material from *Asylums: Essays on the Social Situation of Mental Patients and Other Inmates* by Erving Goffman. Copyright © 1961 by Erving Goffman. Used by permission of Doubleday, an imprint of Knopf Doubleday Publishing Group, a division of Penguin Random House LLC. All rights reserved.

Penguin Random House UK, for material from Goffman, E. (1968) *Asylums: Essays on the Social Situation of Mental Patients and Other Inmates*. Harmondsworth: Penguin Books. Copyright © Erving Goffman, 1961.

QSR International Pty, Ltd, for screenshot reproduced with permission of NVivo qualitative data analysis Software; QSR International Pty Ltd. Version 10, 2012.

Research Council of Norway, for words from Gorard, S. (2012) Mixed methods research in education: some challenges and problems. In Research Council of Norway (ed.) *Mixed Methods in Educational Research: Report on the March Seminar, 2012*, pp. 5–13. Available from: www.uv.uio.no/ils/personer/vit/kirstik/publikasjoner-pdf-filer/klette.-mixed-methods.pdf.

Sage Publications Inc., for material from Patton, M. Q. (1980) *Qualitative Evaluation Methods*, Beverly Hills, CA: Sage; Lee, R. M. (1993) *Doing Research on Sensitive Topics*. London: Sage; Denshire, A. (2014) On auto-ethnography. *Current Sociology Review*, 62 (6), pp. 831–50.

Sheffield Hallam University, Institute of Education, for words from Sheffield Hallam University (2016) *Can Randomised Controlled Trials Revolutionise Educational Research?* Sheffield, UK: Sheffield Institute of Education, Sheffield Hallam University. Available from: www4.shu.ac.uk/research/ceir/randomised-controlled-trials-1.

Springer, for Leech, N. L. and Onwuegbuzie, A. J. (2009) A typology of mixed methods research designs. *Quantity and Quality*, 43 (2), pp. 265–75; Lather, P. (1986) Issues of validity in openly ideological research. *Interchange*, 17 (4), pp. 63–84; Pearce, W. and Raman, S. (2014) The new randomized controlled trials (RCT) movement in public policy: challenges of epistemic governance. *Policy Sciences*, 47 (4), pp. 387–402.

Stanford University Press, for words from Sears, R., Maccoby, E. and Levin, H. (1957) *Patterns of Child Rearing*. Palo Alto, CA: Stanford University Press.

Taylor and Francis (www.tandfonline.com), for Anderson, G. and Arsenault, N. (1998) *Fundamentals of Educational Research* (second edition); Bradley, B. A. and Reinking, D. (2011) Enhancing research and practice in early childhood through formative and design experiments. *Early Child Development and Care*, 181 (3), pp. 305–19; Burgess, R. (ed.) *Issues in Educational Research: Qualitative Methods*; Burgess, R. (ed.) (1993) *Educational Research and Evaluation for Policy and Practice*; Burgess, R. (ed.) (1985) *Issues in Educational Research*; Cuff, E. G. and Payne, G. (1979) *Perspectives in Sociology*; Day, C., Pope, M. and Denicola, P. (eds) (1990) *Insights into Teachers' Thinking and Practice*; Gorard, S. (2002) Fostering scepticism: the importance of warranting claims. *Evaluation and Research in Education*, 16 (3), pp. 136–49; Hammersley, M. (2000) *Taking Sides in Social Research: Essays on Bias and Partisanship*; Hammersley, M. (2015) On ethical principles for social research. *International Journal of Social Research Methodology*, 18 (4), pp. 433–49; Hammersley, M. and Atkinson, P. (1983) *Ethnography: Principles and Practice*; Hitchcock, G. and Hughes, D. (1995) *Research and the Teacher* (second edition); Hong, E., Mason, E., Peng, Y. and Lee, N. (2015) Effects of homework motivation and worry anxiety on homework achievement in

mathematics and English. *Educational Research and Evaluation*, 21 (7–8), pp. 491–514; Morrison, K. R. B. (2009) *Causation in Educational Research*; Piggot-Irvine, E., Rowe, W. and Ferkins, L. (2015) Conceptualizing indicator-domains for evaluating action research. *Educational Action Research*, 23 (4), pp. 545–66; Polkinghorne, D. E. (1995) Narrative configuration in qualitative analysis. *International Journal of Qualitative Studies in Education*, *8* (1) pp. 5–23; Powney, J. and M. Watts (1987) *Interviewing in Educational Research*; Rex, J. (1974) *Approaches to Sociology*; Simons, H. and Usher, R. (eds) (2000) *Situated Ethics in Educational Research*; Walford, G. (ed.) (1994) *Researching the Powerful in Education*; Walford, G. (2001) *Doing Qualitative Educational Research: A Personal Guide to the Research Process*; Walford, G. (2012) Researching the powerful in education: a re-assessment of the problems. *International Journal for Research and Method in Education*, 35 (2), pp. 111–18; Zuber-Skerritt, O. (1996) *New Directions in Action Research*.

University of Chicago Press, for words from Whyte, W. F. (1993) *Street Corner Society*. Chicago, IL: University of Chicago Press.

University of Illinois at Urbana-Champaign, for words from Lansing, J. B., Ginsburg, G. P. and Braaten, K. (1961) *An Investigation of Response Error*. Studies in Consumer Savings, No. 2. Urbana, IL: University of Illinois Bureau of Economic and Business Research.

Disclaimer: The publishers have made every effort to contact authors/copyright holders of works reprinted in the eighth edition of *Research Methods in Education*. We would welcome correspondence from those individuals/companies whom we have been unable to trace.

Part 1
The context of educational research

This part introduces readers to different research traditions, with the advice that 'fitness for purpose' must be the guiding principle: different research paradigms for different research purposes. A major message in this part is that the nature and foundations of educational research have witnessed a proliferation of paradigms over time. From the earlier days of either quantitative or qualitative research have arisen the several approaches introduced here.

This part commences by introducing positivist and scientific contexts of research and some strengths and weaknesses of these for educational research, followed by post-positivist views of research. As an alternative paradigm, the cluster of approaches that can loosely be termed interpretive, naturalistic, phenomenological, interactionist and ethnographic are brought together, and their strengths and weaknesses for educational research are examined. Postmodernist and post-structuralist approaches are also introduced, and these lead into an introduction to complexity theory in educational research. The paradigm of mixed methods research is introduced, and its foundations, strengths, weaknesses, contribution to and practices in educational research are discussed.

Critical theory as a paradigm of educational research is discussed, and its implications for the research are indicated in several ways, resonating with curriculum research, participatory research, feminist research, post-colonial research and queer theory. These are concerned not only with understanding a situation or phenomenon but with *changing* it, often with an explicit political agenda. Critical theory links the conduct of educational research with politics and policy making, and this is reflected in the discussions of research and evaluation, noting how some educational research has become evaluative in nature.

This part includes a new chapter on the role of theory in educational research, indicating its several meanings, its origins and roles in educational research, and what makes a theory interesting and useful. It also includes the discussion of causation in educational research and key elements in understanding and working with causation.

The term *research* itself has many meanings. We restrict its usages here to those activities and undertakings aimed at developing a science of behaviour, the word *science* itself implying both normative and interpretive perspectives. Accordingly, when we speak of social research, we have in mind the systematic and scholarly application of the principles of a science of behaviour to the problems of people within their social contexts, and when we use the term educational research, we likewise have in mind the application of these same principles to the problems of teaching and learning within education and to the clarification of issues having direct or indirect bearing on these concepts.

The nature of enquiry

CHAPTER 1

Setting the field

This large chapter explores the context of educational research. It sets out several foundations on which different kinds of empirical research are constructed:

- the search for understanding
- paradigms of educational research
- scientific and positivistic methodologies
- naturalistic and interpretive methodologies
- post-positivism, post-structuralism and postmodernism
- complexity theory in educational research

Educational researchers cannot simply 'read off' the planning and conduct of research as though one were reading a recipe for baking a cake. Nor is the planning and conduct of research the laboratory world or the field study of the natural scientist. Rather, it is to some degree an art, an iterative and often negotiated process and one in which there are typically trade-offs between what one would like to do and what is actually possible. This book is built on that basis: educational research, far from being a mechanistic exercise, is a deliberative, complex, subtle, challenging, thoughtful activity and often a messier process than researchers would like it to be. This book provides some tools for such deliberation and planning, and hopefully some answers, but beyond that it is for the researcher to consider how to approach, plan, conduct, validate and evaluate the research, how to develop and test theory, how to study and investigate educational matters, how to balance competing demands on the research, and so on. There is no one best way to plan and conduct research, just as there is no one single 'truth' to be discovered. Life is not that easy, unidimensional or straightforwardly understood, just as there are no simple dichotomies in educational research (e.g. quantitative *or* qualitative, objective *or* subjective). Rather, we live in a pluralistic world with many purposes and kinds of research, many realities and lived experiences to catch, many outcomes, theories and explanations, many discoveries to be made, and many considerations and often contradictions or sensitivities to be addressed in the planning and conduct of the research.

Whilst arguing against simple foundationalism, this chapter sets out some conceptions of research which researchers may find helpful in characterizing and deliberating about their studies. The chapter considers paradigms and their possible contribution to educational research, positivism, post-positivism, post-structuralism, postmodernism and interpretive approaches.

1.1 Introduction

Our analysis takes an important notion from Hitchcock and Hughes (1995, p. 21), who suggest that ontological assumptions (assumptions about the nature of reality and the nature of things) give rise to epistemological assumptions (ways of researching and enquiring into the nature of reality and the nature of things); these, in turn, give rise to methodological considerations; and these, in turn, give rise to issues of instrumentation and data collection. Added to ontology and epistemology is axiology (the values and beliefs that we hold). This view moves us beyond regarding research methods as simply a technical exercise to being concerned with understanding the world; this is informed by how we view our world(s), what we take understanding to be, what we see as the purposes of understanding and what is deemed valuable.

1.2 The search for understanding

People have long been concerned to come to grips with their environment and to understand the nature of the phenomena it presents to their senses. The means by which they set out to achieve these ends may be classified into three broad categories: *experience*, *reasoning* and *research* (Mouly, 1978). Far from being independent and mutually exclusive, however, these categories are complementary and overlapping, features most readily in evidence where solutions to complex problems are sought.

In our endeavours to come to terms with day-to-day living, we are heavily dependent upon experience and authority. However, as tools for uncovering ultimate truth, they have limitations. The limitations of personal

experience in the form of common-sense knowing, for instance, can quickly be exposed when compared with features of the scientific approach to problem solving. Consider, for example, the striking differences in the way in which theories are used. Laypeople base them on haphazard events and use them in a loose and uncritical manner. When they are required to test them, they do so in a selective fashion, often choosing only that evidence which is consistent with their hunches and ignoring that which is counter to them. Scientists, by contrast, construct their theories carefully and systematically. Whatever hypotheses they formulate have to be tested empirically so that their explanations have a firm basis in fact. And there is the concept of control distinguishing the layperson's and the scientist's attitude to experience. Laypeople may make little or no attempt to control any extraneous sources of influence when trying to explain an occurrence. Scientists, on the other hand, only too conscious of the multiplicity of causes for a given occurrence, adopt definite techniques and procedures to isolate and test the effect of one or more of the alleged causes. Finally, there is the difference of attitude to the relationships among phenomena. Laypeople's concerns with such relationships may be loose, unsystematic and uncontrolled; the chance occurrence of two events in close proximity is sufficient reason to predicate a causal link between them. Scientists, however, display a much more serious professional concern with relationships and only as a result of rigorous experimentation, investigation and testing will they postulate a relationship between two phenomena.

People attempt to comprehend the world around them by using three types of reasoning: *deductive reasoning*, *inductive reasoning* and the *combined inductive-deductive* approach. Deductive reasoning is based on the syllogism, which was Aristotle's great contribution to formal logic. In its simplest form the syllogism consists of a major premise based on an a priori or self-evident proposition, a minor premise providing a particular instance, and a conclusion. Thus:

All planets orbit the sun;
The earth is a planet;
Therefore the earth orbits the sun.

The assumption underlying the syllogism is that through a sequence of formal steps of logic, from the general to the particular, a valid conclusion can be deduced from a valid premise. Its chief limitation is that it can handle only certain kinds of statement. The syllogism formed the basis of systematic reasoning from the time of its inception until the Renaissance. Thereafter its effectiveness was diminished because it was no longer related to observation and experience and became merely a mental exercise. One of the consequences of this was that empirical evidence as the basis of proof was superseded by authority and the more authorities one could quote, the stronger one's position became.

The history of reasoning was to undergo a dramatic change in the 1600s when Francis Bacon began to lay increasing stress on the observational basis of science. Being critical of the model of deductive reasoning on the grounds that its major premises were often preconceived notions which inevitably bias the conclusions, he proposed in its place the method of inductive reasoning by means of which the study of a number of individual cases would lead to a hypothesis and eventually to a generalization. Mouly (1978) explains it by suggesting that Bacon's basic premise was that, with sufficient data, even if one does not have a preconceived idea of their significance or meaning, nevertheless important relationships and laws will be discovered by the alert observer.

Of course, there are limits to induction as the accumulation of a series of examples does not *prove* a theory; it only *supports* it. Just because all the swans that I have ever seen are white, it does not prove a theory that all swans are white – one day I might come across a black swan, and my theory is destroyed. Induction places limits on prediction. Discoveries of associations of regularities and frequent repetitions may have limited predictive value. We are reminded of Bertrand Russell's (1959) story of the chicken who observed that he was fed each day by the same man, and, because this had happened every day, it would continue to happen, i.e. the chicken had a theory of being fed, but, as Russell remarks, 'the man who has fed the chicken every day throughout its life at last wrings its neck instead' (p. 35), indicating the limits of prediction based on observation. Or, to put it more formally, theory is underdetermined by empirical evidence (Phillips and Burbules, 2000, p. 17). Indeed Popper (1980) notes that the essence of science, what makes a science a science, is the inherent falsifiability of the propositions (in contrast to the views of the method of science as being one of verifiability, as held by logical positivists).

This is not to discard induction: it is often the starting point for science. Rather, it is to caution against assuming that it 'proves' anything. Bacon's major contribution to science was that he was able to rescue it from the stranglehold of the deductive method whose abuse had brought scientific progress to a standstill. He thus directed the attention of scientists to nature for solutions to people's problems, demanding empirical evidence for verification. Logic and authority in themselves were no longer regarded as conclusive means of

proof and instead became sources of hypotheses about the world and its phenomena.

Bacon's inductive method was eventually followed by the inductive-deductive approach which combines Aristotelian deduction with Baconian induction. Here the researcher is involved in a back-and-forth process of induction (from observation to hypothesis, from the specific to the general) and deduction (from hypothesis to implications) (Mouly, 1978). Hypotheses are tested rigorously and, if necessary, revised.

Although both deduction and induction have their weaknesses, their contributions to the development of science are enormous, for example: (1) the suggestion of hypotheses; (2) the logical development of these hypotheses; and (3) the clarification and interpretation of scientific findings and their synthesis into a conceptual framework.

A further means by which we set out to discover truth is *research*. This has been defined by Kerlinger (1970) as the systematic, controlled, empirical and critical investigation of hypothetical propositions about the presumed relations among natural phenomena. Research has three characteristics in particular, which distinguish it from the first means of problem solving identified earlier, namely, experience. First, whereas experience deals with events occurring in a haphazard manner, research is systematic and controlled, basing its operations on the inductive-deductive model outlined above. Second, research is empirical. The scientist turns to experience for validation. As Kerlinger puts it, subjective, personal belief must have a reality check against objective, empirical facts and tests. And third, research is self-correcting. Not only does the scientific method have built-in mechanisms to protect scientists from error as far as is humanly possible, but also their procedures and results are open to public scrutiny by fellow professionals. Incorrect results in time will be found and either revised or discarded (Mouly, 1978). Research is a combination of both experience and reasoning and, as far as the natural sciences are concerned, is to be regarded as the most successful approach to the discovery of truth (Borg, 1963).[1]

1.3 Conceptions of social reality

The views of social science that we have mentioned represent strikingly different ways of looking at social reality and are constructed on correspondingly different ways of interpreting it. We can perhaps most profitably approach these conceptions of the social world by examining the explicit and implicit assumptions underpinning them. Our analysis is based on the work of Burrell and Morgan (1979), who identified four sets of such assumptions.

First, there are assumptions of an ontological kind – assumptions which concern the very nature or essence of the social phenomena being investigated. Thus, the authors ask, is social reality external to individuals – imposing itself on their consciousness from without – or is it the product of individual consciousness? Is reality of an objective nature, or the result of individual cognition? Is it a given 'out there' in the world, or is it created by one's own mind? Is there a world which exists independent of the individual and which the researcher can observe, discovering relationships, regularities, causal explanations, and which can be tested empirically and repeatedly (i.e. under similar conditions) (cf. Pring, 2015, p. 64)? These questions spring directly from what philosophy terms the nominalist–realist debate. The former view holds that objects of thought are merely words and that there is no independently accessible thing constituting the meaning of a word. The realist position, however, contends that objects have an independent existence and are not dependent for it on the knower. The fact that I can see a dog is not simply because of my perception or cognition but because a dog exists independent of me.

The second set of assumptions identified by Burrell and Morgan are of an epistemological kind. These concern the very bases of knowledge – its nature and forms, how it can be acquired and how communicated to other human beings. How one aligns oneself in this particular debate profoundly affects how one will go about uncovering knowledge of social behaviour. The view that knowledge is hard, objective and tangible will demand of researchers an observer role, together with an allegiance to the methods of natural science; to see knowledge as personal, subjective and unique, however, imposes on researchers an involvement with their subjects and a rejection of the ways of the natural scientist. To subscribe to the former is to be positivist; to the latter, anti-positivist or post-positivist.

The third set of assumptions concern human nature and, in particular, the relationship between human beings and their environment. Since the human being is both its subject and object of study, the consequences for social science of assumptions of this kind are far-reaching. Two images of human beings emerge from such assumptions – the one portrays them as responding mechanically and deterministically to their environment, i.e. as products of the environment, controlled like puppets; the other, as initiators of their own actions with free will and creativity, producing their own environments. The difference is between *determinism* and *voluntarism* respectively (Burrell and Morgan, 1979), between *structure* and *agency*. Human action involves

some combination of these two, polarized here for the sake of conceptual clarity.

It follows from what we have said so far that the three sets of assumptions identified above have direct implications for the methodological concerns of researchers, since the contrasting ontologies, epistemologies and models of human beings will, in turn, suggest different research methods. Investigators adopting an objectivist (or positivist) approach to the social world and who treat it like the world of natural phenomena as being real and external to the individual will choose from a range of options such as surveys, experiments and the like. Others favouring the more subjectivist (or anti-positivist) approach and who view the social world as being of a much more personal and humanly created kind will select from a comparable range of recent and emerging techniques – accounts, participant observation, interpretive approaches and personal constructs, for example.

Where one subscribes to the view which treats the social world like the natural world – as if it were an external and objective reality – then scientific investigation will be directed at analysing the relationships and regularities between selected factors in that world. It will be concerned with identifying and defining elements and discovering ways in which their relationships can be expressed. Hence, methodological issues, of fundamental importance, are thus the concepts themselves, their measurement and the identification of underlying themes in a search for universal laws which explain and govern that which is being observed (Burrell and Morgan, 1979). An approach characterized by procedures and methods designed to discover general laws may be referred to as *nomothetic*. Here is not the place to debate whether social life is 'law-like' (i.e. can be explained by universal laws) in the same way as that mooted in the natural sciences (but see Kincaid, 2004) or whether social life is quintessentially different from the natural sciences such that 'law-like' accounts are simply a search for the impossible and untenable.

However, if one favours the alternative view of social reality which stresses the importance of the subjective experience of individuals in the creation of the social world, then the search for understanding focuses upon different issues and approaches them in different ways. The principal concern is with an *understanding* of the way in which individuals and social groups create, modify and interpret the world in which they find themselves. As Burrell and Morgan (1979) observe, emphasis here is placed on explanation and understanding of the unique and the particular individual cases (however defined: see Chapter 19 on case study, in which emphasis is placed on the denotation of what is the case: an individual, a group, a class, an institution etc.) rather than the general and the universal. In its emphasis on the particular and individual case, this approach to understanding individual (however defined) behaviour may be termed *idiographic*.

In this review of Burrell and Morgan's analysis of the ontological, epistemological, human and methodological assumptions underlying two ways of conceiving social reality, we have laid the foundations for a more extended study of the two contrasting perspectives evident in the practices of researchers investigating human behaviour and, by adoption, educational problems. Figure 1.1 summarizes these assumptions along a subjective/objective dimension. It identifies the four

A scheme for analysing assumptions about the nature of social science

The subjectivist approach to social science				The objectivist approach to social science
Nominalism	←	Ontology	→	Realism
Anti-positivism	←	Epistemology	→	Positivism
Voluntarism	←	Human nature	→	Determinism
Idiographic	←	Methodology	→	Nomothetic

FIGURE 1.1 The subjective-objective dimension
Source: Burrell and Morgan (1979)

sets of assumptions by using terms we have adopted in the text and by which they are known in the literature of social philosophy.

Each of the two perspectives on the study of human behaviour outlined above has profound implications for research in classrooms and schools. The choice of problem, the formulation of questions to be answered, the characterization of students and teachers, methodological concerns, the kinds of data sought and their mode of treatment, all are influenced by the viewpoint held. Some idea of the considerable practical implications of the contrasting views can be gained by examining Table 1.1, which compares them with respect to a number of critical issues within a broadly societal and

TABLE 1.1 ALTERNATIVE BASES FOR INTERPRETING SOCIAL REALITY

	Conceptions of social reality	
Dimensions of comparison	*Objectivist*	*Subjectivist*
Philosophical basis	Realism: the world exists and is knowable as it really is. Organizations are real entities with a life of their own.	Idealism: the world exists but different people construe it in very different ways. Organizations are invented social reality.
The role of social science	Discovering the universal laws of society and human conduct within it.	Discovering how different people interpret the world in which they live.
Basic units of social reality	The collectivity: society or organizations.	Individuals acting singly or together.
Methods of understanding	Identifying conditions or relationships which permit the collectivity to exist. Conceiving what these conditions and relationships are.	Interpretation of the subjective meanings which individuals place upon their action. Discovering the subjective rules for such action.
Theory	A rational edifice built by scientists to explain human behaviour.	Sets of meanings which people use to make sense of their world and behaviour within it.
Research	Experimental or quasi-experimental validation of theory.	The search for meaningful relationships and the discovery of their consequences for action.
Methodology	Abstraction of reality, especially through mathematical models and quantitative analysis.	The representation of reality for purposes of comparison. Analysis of language and meaning.
Society	Ordered. Governed by a uniform set of values and made possible only by those values.	Conflicted. Governed by the values of people with access to power.
Organizations	Goal oriented. Independent of people. Instruments of order in society serving both society and the individual.	Dependent upon people and their goals. Instruments of power which some people control and can use to attain ends which seem good to them.
Organizational pathologies	Organizations get out of kilter with social values and individual needs.	Given diverse human ends, there is always conflict among people acting to pursue them.
Prescription for change	Change the structure of the organization to meet social values and individual needs.	Find out what values are embodied in organizational action and whose they are. Change the people or change their values if you can.

Source: Adapted from Barr Greenfield (1975)

organizational framework. Implications of the two perspectives for educational research unfolds in the course of the text.

1.4 Paradigms

Educational research has absorbed several competing views of the social sciences – the scientific view and an interpretive view – and several others that we explore in this book, including critical theory and feminist theory. Some views hold that the social sciences are essentially the same as the natural sciences and are therefore concerned with discovering natural and universal laws regulating and determining individual and social behaviour. The interpretive view, however, while sharing the rigour of the natural sciences and the concern of social science to describe and explain human behaviour, emphasizes how people differ from inanimate natural phenomena and, indeed, from each other. These contending views – and also their corresponding reflections in educational research – stem in the first instance from different conceptions of social realities and of individual and social behaviour. We examine these in a little more detail.

Since the groundbreaking work of Kuhn (1962), approaches to methodology in research have been informed by discussions of 'paradigms' and communities of scholars. A paradigm is a way of looking at or researching phenomena, a world view, a view of what counts as accepted or correct scientific knowledge or way of working, an 'accepted model or pattern' (Kuhn, 1962, p. 23), a shared belief system or set of principles, the identity of a research community, a way of pursuing knowledge, consensus on what problems are to be investigated and how to investigate them, typical solutions to problems, and an understanding that is more acceptable than its rivals.

A notable example of this is the old paradigm that placed the Earth at the centre of the universe, only to be replaced by the Copernican heliocentric model, as evidence and explanation became more persuasive of the new paradigm. Importantly, one has to note that the old orthodoxy retained its value for generations because it was supported by respected and powerful scientists and, indeed, others (witness the attempts made by the Catholic Church to silence Galileo in his advocacy of the heliocentric model of the universe). Another example is where the Newtonian view of the mechanical universe has been replaced by the Einsteinian view of a relativistic, evolving universe. More recently still, the idea of a value-free, neutral, objective, positivist science has been replaced by a post-positivist, critical realist view of science with its hallmarks of conjecture and refutation (Popper, 1980) and with the ability for falsification being the distinguishing feature of science. Further, social science has recognized the importance of the (subjective) value systems of researchers, phenomenology, subjectivity, the need for reflexivity in research (discussed later in this book), the value of qualitative and mixed methods approaches to research, and the contribution of critical theory and feminist approaches to research methodologies and principles.

Paradigms are not simply methodologies (Hammersley, 2013, p. 15); they are ways of looking at the world, different assumptions about what the world is like and how we can understand or know about it. This raises the question of whether paradigms can live together, whether they are compatible or, since they constitute fundamentally different ways of looking at the world, they are incommensurate (which raises questions for mixed methods research – see Chapter 2). At issue here is the significance of regarding approaches to research as underpinned by different paradigms, an important characteristic of which is their incommensurability with each other (i.e. one cannot hold two distinct paradigms simultaneously as there are no common principles, standards or measures).

As more knowledge is acquired to challenge an existing paradigm, such that the original paradigm cannot explain a phenomenon as well as the new paradigm, there comes about a 'scientific revolution', a paradigm shift, in which the new paradigm replaces the old as the orthodoxy – the 'normal science' – of the day. Kuhn's (1962) notions of paradigms and paradigm shifts link here objects of study and communities of scholars, where the field of knowledge or paradigm is seen to be only as good as the evidence and the respect in which it is held by 'authorities'.

Part 1 sets out several paradigms of educational research and these are introduced in Chapters 1 to 3.

Social science research is marked by paradigmatic pluralism and multiple ways of construing paradigms. For example, Pring (2015) contrasts two paradigms (pp. 63–74). The first paradigm espouses the view that there is an objective reality which exists independent of the individual and comprises causally interacting elements which are available for observation; that different sciences (e.g. social, physical) can be used to define that reality once consensus has been reached on what that objective reality is; that the research is replicable and cumulative, i.e. a scientifically rooted body of knowledge can be gathered and checked for correspondence to the world as it is (the correspondence theory of truth) (pp. 63–4). Such a view resonates with Hammersley's (2013) summary of quantitative research which is characterized by hypothesis testing, numerical

data, 'procedural objectivity', generalization, the identification of 'systematic patterns of association' and the isolation and control of variables (pp. 10–11).

The second paradigm, by contrast, espouses the view that the world consists of ideas, i.e. a social construction, and that researchers are part of the world which they are researching, that meanings are negotiated between participants (including the researcher), that an objective test of truth is replaced by a consensus theory of truth, that ideas of the world do not exist independently of those who hold them (i.e. require a redefinition of 'objective' and 'subjective'), that multiple realities exist and that what is being researched is context-specific (Pring, 2015, pp. 65–6). Such a view accords with Hammersley's definition of qualitative research as that which uses less structured data, which emphasizes the central place of subjectivity in the research process and which studies 'a small number of naturally occurring cases in detail' using verbal rather than statistical analysis (Hammersley, 2013, p. 12).

However, Pring's (2015) point is not simply to set out these two paradigms, but to argue that they constitute a false dualism that should be rejected, as they artificially compel the researcher to make an either/or choice of paradigms and, thereby, misrepresent the world as multiply meaningful and both independent of and part of the researcher, not only a social construction. He argues (p. 69) that, just as an independent physical world must exist in order for researchers to construe it, the same can be said of the social world – there must be independent actors and social worlds in order for apperception and social construction of it to make sense.

Pring cautions against adopting a priori either a quantitative or qualitative view of the world as this massively over-simplifies the real world, which is complex and complicated. Rather, how we pursue the research depends on what the research is about, and this recognizes that social constructions vary from social group to social group and humans can be both the object and subject of research (2015, p. 73).

Pring is not alone in characterizing different paradigms of educational research. For example, Creswell (2013) notes four 'philosophical worldviews' (pp. 7ff.): post-positivism, constructivism, advocacy/participatory and pragmatism. These are discussed in Chapters 2 and 3. Here we note that the advocacy/participatory paradigm concerns the disempowered and marginalized, and it studies oppression and lack of voice; this brings it under the umbrella of critical approaches which we discuss in Chapter 3, including gender, race, ethnicity, disability, sexual orientation, socio-economic status and differentials of power that prop up inequality.

Lather (2004) sets out four paradigms: prediction (positivism); understanding (interpretive approaches); emancipatory (critical theoretical approaches); and deconstruction (post-structuralist). We discuss these in Chapters 1 to 3. Lukenchuk (2013) identifies six paradigms which, she notes, are not exhaustive (pp. 66ff.):

- Empirical-analytic (empiricist; scientific; concerned with prediction and control; quantitative; experimental; correlational; causal; explanatory; probabilistic; fallibilistic; concerned with warrants for knowledge claims; quantitative);
- Pragmatic (focus on 'what works'; trial and error; problem-centred; practical; experimental; action oriented; utility oriented; practitioner research; qualitative and quantitative);
- Interpretive (hermeneutic and existential understanding; meaning-making; phenomenological; qualitative; naturalistic; constructivist; interactionist; verstehen approaches; ethnographic; qualitative);
- Critical (ideology-critical; concerned with analysis of power and ideology; consciousness-raising; emancipatory and concerned with advocacy/participatory approaches; transformatory; politically oriented and activist; qualitative and quantitative);
- Post-structuralist (anti-foundation knowledge; deconstructionist; interpretation of life as discourse and texts; transformative; qualitative);
- Transcendental (asserts reason, intuition, mysticism, revelation as ways of knowing: mind, body, soul and spirit; life as directed by an 'internal moral compass'; foundational; qualitative).

This is not to say that paradigms necessarily *drive* the research, as research is driven by the purposes of the research. Indeed we can ask whether we need paradigmatic thinking at all in order to do research. Rather, it is to say that the purposes and nature of the research may be clarified by drawing on one or more of these paradigms; the paradigms can clarify and organize the thinking about the research. Further, it is not to say that these paradigms each have an undisputed coherence, unity or unproblematic singularity of conception. Rather, they are characterizations, ideal types, typifications and simplifications for ease of initial understanding, recognizing that this blurs the many variations that lie within each of them, and, indeed, may overlook the overlaps between them; each paradigm is not all of a single type and they are by no means mutually exclusive. To consider them as mutually exclusive is to prolong the unnecessary 'paradigm wars' to which Gage (1989) alluded so compellingly.

Because of its significance for the epistemological basis of social science and its consequences for educational research, we devote discussion in this chapter to the debate on positivism and anti-positivism/post-positivism, and on alternative paradigms and rationales for understanding educational research.

1.5 Positivism

Although positivism has been a recurrent theme in the history of western thought from the Ancient Greeks to the present, it is historically associated with the nineteenth-century French philosopher, Auguste Comte, who was the first thinker to use the word for a philosophical position (Beck, 1979) and who gave rise to sociology as a distinct discipline. His positivism turns to observation and reason as means of understanding behaviour, i.e. empirical observation and verification; explanation proceeds by way of scientific description. In his study of the history of the philosophy and methodology of science, Oldroyd (1986) says that, in this view, social phenomena could be researched in ways similar to natural, physical phenomena, i.e. generating laws and theories that could be investigated empirically.

Comte's position was to lead to a general doctrine of positivism which held that all genuine knowledge is based on sensory experience and can only be advanced by means of observation and experiment: the scientific method. Following in the empiricist tradition, it limited enquiry and belief to what can be firmly established and in thus abandoning metaphysical and speculative attempts to gain knowledge by reason alone, the movement developed a rigorous orientation to social facts and natural phenomena to be investigated empirically (Beck, 1979). Taking account of this, matters of values were out of court for the positivist, as they were not susceptible to observation evidence, i.e. there is a separation between facts and values.

With its emphasis on observational evidence and the scientific method, positivism accords significance to sensory experience (empiricism), observational description (e.g. ruling our inferences about actors' intentions, thoughts or attitudes), operationalism, 'methodical control', measurement, hypothesis testing and replicability through the specification of explicit and transparent procedures for conducting research (Hammersley, 2013, pp. 23–4). Hammersley notes that the terms 'positivism' and 'empiricism' are often regarded as synonymous with each other (p. 23), but to equate positivism simply with quantitative approaches is misguided, as qualitative data are equally well embraced within empiricism. Indeed he notes that ethnographers and discourse analysts rely on careful observational data (pp. 24–5).

Though the term positivism is used by philosophers and social scientists, a residual meaning derives from an acceptance of natural science as the paradigm of human knowledge (Duncan, 1968). This includes the following connected suppositions, identified by Giddens (1975). First, the methodological procedures of natural science may be directly applied to the social sciences. Positivism here implies a particular stance concerning the social scientist as an observer of social reality. Second, the end-product of investigations by social scientists can be formulated in terms parallel to those of natural science. This means that their analyses must be expressed in laws or law-like generalizations of the same kind that have been established in relation to natural phenomena. Positivism claims that science provides us with the clearest possible ideal of knowledge.

Where positivism is less successful, however, is in its application to the study of human behaviour, where the immense complexity of human nature and the elusive and intangible quality of social phenomena contrast strikingly with the order and regularity of the natural world. This point is apparent in the contexts of classrooms and schools where the problems of teaching, learning and human interaction present the positivistic researcher with a mammoth challenge.

We now look more closely at some of the features of the scientific method that is underpinned by positivism.

1.6 The assumptions and nature of science

We begin with an examination of the tenets of scientific faith: the kinds of assumptions held by scientists, often implicitly, as they go about their daily work. First, there is the assumption of *determinism*. This means simply that events have causes; that events are determined by other circumstances; and science proceeds on the belief that these causal links can eventually be uncovered and understood. Moreover, not only are events in the natural world determined by other circumstances, but there is regularity about the way in which they are determined: the universe does not behave capriciously. It is the ultimate aim of scientists to formulate laws to account for the happenings in the world, thus giving them a firm basis for prediction and control.

The second assumption is that of *empiricism*, which holds that certain kinds of reliable knowledge can only derive from experience. This is an example of foundationalism. In this case, to quote the philosopher John

Locke (1959): 'whence has it [the mind] all the materials of reason and knowledge? To this, I answer, in one word, from experience. In that all knowledge is founded; and from that it ultimately derives itself' (p. 26). Experience means sensory experience, and this contrasts with the rationalist epistemology in which reason rules supreme. In empiricism, experience alone provides the warrant for, or justification of, a knowledge claim, which is brought to the scientific community for acceptance. Such empiricism gives rise to the need for the operationalization of concepts, for example, creativity, intelligence, ability (Phillips and Burbules, 2000, p. 10), in order for them to be observable. Empiricism (and positivism) does not preclude non-experimental studies, nor does it prescribe only quantitative research.

In practice, empiricism means scientifically that the tenability of a theory or hypothesis depends on the nature of the empirical evidence for its support. 'Empirical' here means that which is verifiable by observation, direct experience and evidence, data-yielding proof or strong confirmation, in probability terms, of a theory or hypothesis in a research setting.

Mouly (1978) identifies five steps in the process of empirical science:

1. *Experience* – the starting point of scientific endeavour at the most elementary level;
2. *Classification* – the formal systematization of otherwise incomprehensible masses of data;
3. *Quantification* – a more sophisticated stage where precision of measurement allows more adequate analysis of phenomena by mathematical means;
4. *Discovery of relationships* – the identification and classification of functional relationships among phenomena;
5. *Approximation to the truth* – science proceeds by gradual approximation to the truth.

The third assumption underlying the work of the scientist is the principle of *parsimony*. The basic idea is that phenomena should be explained in the most economical way possible. As Einstein was known to remark, one should make matters as simple as possible, but no simpler! The first historical statement of the principle was by William of Occam when he said that explanatory principles (entities) should not be needlessly multiplied ('Occam's razor'), i.e. that it is preferable to account for a phenomenon by two concepts rather than three; that a simple theory is to be preferred to a complex one.

The final assumption, that of *generality*, played an important part in both the deductive and inductive methods of reasoning. Indeed, historically speaking, it was the problematic relationship between the concrete particular and the abstract general that was to result in two competing theories of knowledge – the rational and the empirical. Beginning with observations of the particular, scientists set out to generalize their findings to the world at large. This is because they are concerned ultimately with explanation. Of course, the concept of generality presents much less of a problem to natural scientists working chiefly with inanimate matter than to human scientists who, of necessity having to deal with samples of larger human populations, must exercise great caution when generalizing their findings to the particular parent populations.

We come now to the core question: What is science? Kerlinger (1970) points out that in the scientific world itself two broad views of science may be found: the *static* and the *dynamic*. The *static* view, which has particular appeal for laypeople, is that science is an activity that contributes systematized information to the world. The work of the scientist is to uncover new facts and add them to the existing corpus of knowledge. Science is thus seen as an accumulated body of

BOX 1.1 THE FUNCTIONS OF SCIENCE

1. Its problem-seeking, question-asking, hunch-encouraging, hypotheses-producing function.
2. Its testing, checking, certifying function; its trying out and testing of hypotheses; its repetition and checking of experiments; its piling up of facts.
3. Its organizing, theorizing, structuring function; its search for larger and larger generalizations.
4. Its history-collecting, scholarly function.
5. Its technological side; instruments, methods, techniques.
6. Its administrative, executive and organizational side.
7. Its publicizing and educational functions.
8. Its applications to human use.
9. Its appreciation, enjoyment, celebration and glorification.

Source: Maslow (1954)

findings, the emphasis being chiefly on the present state of knowledge and adding to it.[2] The *dynamic* view, by contrast, conceives science more as an activity, as something that scientists *do*. According to this conception it is important to have an accumulated body of knowledge of course, but what really matter most are the discoveries that scientists make. The emphasis here, then, is more on the heuristic nature of science.

Contrasting views exist on the functions of science. We give a composite summary of these in Box 1.1. For professional scientists, however, science is seen as a way of comprehending the world; as a means of explanation and understanding, of prediction and control. For them the ultimate aim of science is theory, and we discuss this in Chapter 4.

We look now in more detail at two such tools which play a crucial role in science – the concept and the hypothesis.

1.7 The tools of science

Concepts express generalizations from particulars – anger, achievement, alienation, velocity, intelligence, democracy. Examining these examples more closely, we see that each is a word representing an idea: more accurately, a concept is the relationship between the word (or symbol) and an idea or conception. Whoever we are and whatever we do, we all make use of concepts. Naturally, some are shared and used by all groups of people within the same culture – child, love, justice, for example; others, however, have a restricted currency and are used only by certain groups, specialists or members of professions – idioglossia, retroactive inhibition, anticipatory socialization.

Concepts enable us to impose some sort of meaning on the world; through them reality is given sense, order and coherence. They are the means by which we are able to come to terms with our experience. How we perceive the world, then, is highly dependent on the repertoire of concepts that we have. The more we have, the more sense data we can pick up and the surer will be our perceptual (and cognitive) grasp of whatever is 'out there'. If our perceptions of the world are determined by the concepts available to us, it follows that people with differing sets of concepts will tend to view the 'same' objective reality differently – a doctor diagnosing an illness will draw upon a vastly different range of concepts from, say, the restricted and perhaps simplistic notions of the layperson in that context.

So where is all this leading? Simply to this: social scientists have likewise developed, or appropriated by giving precise meaning to, a set of concepts which enable them to shape their perceptions of the world in a particular way, to represent that slice of reality which is their special study. And collectively, these concepts form part of their wider meaning system which permits them to give accounts of that reality, accounts which are rooted and validated in the direct experience of everyday life, for example, the concept of social class which offers researchers 'a rule, a grid, even though vague at times, to use in talking about certain sorts of experience that have to do with economic position, lifestyle, life-chances, and so on' (Hughes, 1976, p. 34).

There are two important points to stress when considering scientific concepts. The first is that they do not exist independently of us: they are our inventions, enabling us to acquire some understanding of nature. The second is that they are limited in number and in this way contrast with the infinite number of phenomena they are required to explain.

A second tool of great importance to the scientist is the *hypothesis*. It is from this that much research proceeds, especially where cause-and-effect or concomitant relationships are being investigated. The hypothesis has been defined by Kerlinger (1970) as a conjectural statement of the relations between two or more variables, or 'an educated guess', though it is unlike an educated guess in that it is often the result of considerable study, reflective thinking and observation. Medawar (1972) writes of the hypothesis and its function as being speculative and imaginative preconceptions or conjectures about what might be true, which are subject to criticism to see if they really are like the phenomenon in question. As he remarks, scientific reasoning is a dialogue between the 'imaginative and the critical', the 'possible and the actual', between 'what might be true and what is in fact the case' (Medawar, 1972, p. 22).

Kerlinger (1970) has identified two criteria for 'good' hypotheses. The first is that hypotheses are statements about the relations between variables; and second, that hypotheses carry clear implications for testing the stated relations. To these he adds two ancillary criteria: that hypotheses disclose compatibility with current knowledge; and that they are expressed as economically as possible. Thus if we conjecture that social class background determines academic achievement, we have a relationship between one variable, social class, and another, academic achievement. And since both can be measured, the primary criteria specified by Kerlinger can be met. Neither do they violate the ancillary criteria he proposed (see also Box 1.2).

Kerlinger further identifies four reasons for the importance of hypotheses as tools of research. First, they organize the efforts of researchers. The relationship expressed in the hypothesis indicates what they should do. They enable them to understand the problem

BOX 1.2 THE HYPOTHESIS

Once one has a hypothesis to work on, the scientist can move forward; the hypothesis will guide the researcher on the selection of some observations rather than others and will suggest experiments. Scientists soon learn by experience the characteristics of a good hypothesis. A hypothesis that is so loose as to accommodate *any* phenomenon tells us precisely nothing; the more phenomena it prohibits, the more informative it is.

A good hypothesis must also have *logical immediacy*, i.e. it must provide an explanation of whatever it is that needs to be explained and not an explanation of other phenomena. Logical immediacy in a hypothesis means that it can be tested by comparatively direct and practicable means. A large part of the *art of the soluble* is the art of devising hypotheses that can be tested by practicable experiments.

Source: Adapted from Medawar (1981)

with greater clarity and provide them with a framework for collecting, analysing and interpreting their data. Second, they are, in Kerlinger's words, the working instruments of theory. They can be deduced from theory or from other hypotheses. Third, they can be tested, empirically or experimentally, resulting in confirmation or rejection. And there is always the possibility that a hypothesis, once supported and established, may become a law. And fourth, hypotheses are powerful tools for the advancement of knowledge because, as Kerlinger explains, they enable us to get outside ourselves. Hypotheses and concepts play a crucial part in the scientific method and it is to this that we now turn our attention.

1.8 The scientific method

If the most distinctive feature of science is its empirical nature, the next most important characteristic is its set of procedures which show not only how findings have been arrived at, but are sufficiently clear for fellow-scientists to repeat them, i.e. to check them out with the same or other materials and thereby test the results. As Cuff and Payne (1979) say: 'A scientific approach necessarily involves standards and procedures for demonstrating the "empirical warrant" of its findings, showing the match or fit between its statements and what is happening or has happened in the world' (Cuff and Payne, 1979, p. 4). For convenience we will call these standards and procedures 'the scientific method', though this can be somewhat misleading, as the combination of the definite article, adjective and singular noun risks conjuring up a single invariant approach to problem solving. Yet there is much more to it than this. The term in fact cloaks a number of methods which vary in their degree of sophistication depending on their function and the particular stage of development a science has reached.

The scientific method initially involves systematic observation, moving to interconnecting ideas coherently and without internal contradictions (creating a scientific model), which is then tested by further observations (Capra and Luisi, 2014). Box 1.3 sets out the sequence of stages through which a science normally passes in its development or, perhaps more realistically, that are constantly present in its progress and on which scientists may draw depending on the kind of information they seek or the kind of problem confronting them.

BOX 1.3 STAGES IN THE DEVELOPMENT OF A SCIENCE

1. Definition of the science and identification of the phenomena that are to be subsumed under it.
2. Observational stage at which the relevant factors, variables or items are identified and labelled; and at which categories and taxonomies are developed.
3. Correlational research in which variables and parameters are related to one another and information is systematically integrated as theories begin to develop.
4. The systematic and controlled manipulation of variables to see if experiments will produce expected results, thus moving from correlation to causality.
5. The firm establishment of a body of theory as the outcomes of the earlier stages are accumulated. Depending on the nature of the phenomena under scrutiny, laws may be formulated and systematized.
6. The use of the established body of theory in the resolution of problems or as a source of further hypotheses.

Of particular interest in our efforts to elucidate the term 'scientific method' are stages 2, 3 and 4. Stage 2 is a relatively uncomplicated point at which the researcher is content to observe and record facts and possibly arrive at some system of classification. Much research in the field of education is conducted in this way, for example, surveys and case studies. Stage 3 establishes relationships between variables within a loose framework of inchoate theory. Stage 4 is the most sophisticated stage and often the one that many people equate exclusively with the scientific method. In order to arrive at causality, as distinct from mere measures of association, researchers here design experimental situations in which variables are manipulated to test their chosen hypotheses. This process moves from early, inchoate ideas, to more rigorous hypotheses, to empirical testing of those hypotheses, thence to confirmation or modification of the hypotheses (Kerlinger, 1970).

Hitchcock and Hughes (1995, p. 23) suggest an eight-stage model of the scientific method that echoes Kerlinger. This is represented in Box 1.4.

The elements the researchers fasten on to will naturally be suitable for scientific formulation; this means simply that they will possess quantitative aspects. Their principal working tool will be the hypothesis which, as we have seen, is a statement indicating a relationship (or its absence) between two or more of the chosen elements and stated in such a way as to carry clear implications for testing. Researchers then choose the most appropriate method and put their hypotheses to the test.

1.9 Criticisms of positivism and the scientific method

In spite of the scientific enterprise's proven success using positivism – especially in the field of natural science – its ontological and epistemological bases have been the focus of sustained and sometimes vehement criticism from some quarters. Beginning in the second half of the nineteenth century, the revolt against positivism occurred on a broad front. Essentially, it has been a reaction against the world picture projected by science which, it is contended, undermines life and mind. The precise target of the anti-positivists' attack has been science's mechanistic and reductionist view of nature which, by definition, regards life in measurable terms rather than inner experience, and excludes notions of choice, freedom, individuality and moral responsibility, regarding the universe as a living organism rather than as a machine (e.g. Nesfield-Cookson, 1987).

Here the putative objectivity of science is called into question, and objectivity is treated as problematic. Kettley (2012), for example, notes that objective knowledge is often treated as unproblematic and viewed through simplistic, unacceptably reductionist lenses in which empiricism is reduced to knowing through observation, positivism is viewed as Comte's rebuttal of metaphysics, that there is a unity between the scientific method and Durkheim's positivism, and realism is a synonym for undisputed existence (p. 71). However, he contends, objective knowledge is actually contested, subjective meanings affect or refract views of what are generally considered to be objective knowledge and objectivity (e.g. social facts) which do not necessarily reside in the phenomenon itself but in the subjective values of the researcher (p. 72), and that equating the scientific methods with positivism overlooks the important distinction between induction and deduction. Douglas (2004) notes that the very term 'objective' is fraught with definitional problems, and he gives several senses in which it is used, including, for example: manipulable, detached, procedural, value-neutral and value-free.

The point is well made: objectivity and objective knowledge are beset with problems, and researchers are well advised to avoid simple dichotomies or absolutist ideal types: objective or subjective, induction or deduction, quantitative or qualitative. Rather, there is no unified objectivist or subjectivist paradigm (Kettley,

BOX 1.4 AN EIGHT-STAGE MODEL OF THE SCIENTIFIC METHOD

Stage 1: Hypotheses, hunches and guesses
Stage 2: Experiment designed; samples taken; variables isolated
Stage 3: Correlations observed; patterns identified
Stage 4: Hypotheses formed to explain regularities
Stage 5: Explanations and predictions tested; falsifiability
Stage 6: Laws developed or disconfirmation (hypothesis rejected)
Stage 7: Generalizations made
Stage 8: New theories

2012, p. 76); objective reality is constructed subjectively; positivism is not a unified, singular, coherent tenet; hypothesis formation is a human act that derives in part from the subjective views of the researcher (and these subjective views can differ sharply); aggregated data do not override or negate the constructions and meanings accorded to a situation by individuals; and the assumption of linear relationships is frustrated by a non-linear world (pp. 76–7).

Another challenge to the claims of positivism came from Søren Kierkegaard, the Danish philosopher, one of the originators of existentialism. Kierkegaard was concerned with individuals and their need to fulfil themselves to the highest level of development. This realization of a person's potential was for him the meaning of existence which he saw as concrete and individual, unique and irreducible, not amenable to conceptualization (Beck, 1979). Features of the age in which we live – the ascendancy of scientific and technological progress – militate against the achievement of this end and contribute to the dehumanization of the individual. In his desire to free people from their illusions, the illusion Kierkegaard was most concerned about was that of objectivity. By this he meant the imposition of rules of behaviour and thought, and the making of a person into an observer set on discovering general laws governing human behaviour. The capacity for subjectivity, he argued, should be regained and retained. This he regarded as the ability to consider one's own relationship to whatever constitutes the focus of enquiry.

Also concerned with the dehumanizing effects of the social sciences is Ions (1977). While acknowledging that they can take much credit for throwing light in dark corners, he expresses serious concern at the way in which quantification and computation, assisted by statistical theory and method, are used. He argues that quantification is a form of collectivism, but that this runs the risk of depersonalization. His objection is not directed at quantification per se, but at quantification when it becomes an end in itself, replacing humane study which seeks to investigate and shed light on the human condition (Ions, 1977). This echoes Horkheimer's (1972) powerful critique of positivism as the mathematization of concepts about nature and of scientism – science's belief in itself as the only way of conducting research and explaining phenomena.

Another forceful critic of the objective consciousness has been Roszak (1970, 1972), who argues that science, in its pursuit of objectivity, is a form of alienation from our true selves and from nature. The justification for any intellectual activity lies in the effect it has on increasing our awareness and degree of consciousness, but this increase, some claim, has been retarded in our time by the excessive influence that the positivist paradigm has exerted on areas of our intellectual life. Holbrook (1977), for example, affording consciousness a central position in human existence and deeply concerned with what happens to it, condemns positivism and empiricism for their bankruptcy of the inner world, morality and subjectivity.

Hampden-Turner (1970) concludes that the social science view of human beings is a restricted image of humans when social scientists concentrate on the repetitive, predictable and invariant aspects of the person; on 'visible externalities' to the exclusion of the subjective world; and on the parts of the person in their endeavours to understand the whole.

Habermas (1972), in keeping with the Frankfurt School of critical theory (discussed in Chapter 3), provides a corrosive critique of positivism, arguing that the scientific mentality has been elevated to an almost unassailable position – almost to the level of a religion (scientism) – as being the only epistemology of the west. In this view all knowledge becomes equated with scientific knowledge. This neglects hermeneutic, aesthetic, critical, moral, creative and other forms of knowledge. It reduces behaviour to technicism.

Positivism's concern for control and, thereby, its appeal to the passivity of behaviourism and for instrumental reason is a serious danger to the more open-ended, creative, humanitarian aspects of social behaviour. Habermas (1972, 1974) and Horkheimer (1972) argue that scientism silences an important debate about values, informed opinion, moral judgements and beliefs. Scientific explanation seems to be the only means of explaining behaviour, and, for them, this seriously diminishes the very characteristics that make humans human. It makes for a society without conscience. Positivism is unable to answer many interesting or important areas of life (Habermas, 1972, p. 300), resonating with Wittgenstein's (1974) comment that when all possible scientific questions have been addressed, they have left untouched the main problems of life.

Other criticisms are commonly levelled at positivistic social science. One is that it fails to take account of our unique ability to interpret our experiences and represent them to ourselves. How we make sense of the social world resides in our distinctively human nature, and we have to take account of this in recognizing that the social world is not the same as an object of science (Pring, 2015, p. 115) (though Durkheim noted that there are 'social facts', i.e. those that transcend individuals' interpretations and constructions). We can, and do, construct theories about ourselves and our world,

and we act on these theories. In failing to recognize this, positivistic social science is said to ignore the profound differences between itself and the natural sciences. Social science, unlike natural science, stands in a subject–subject rather than a subject–object relation to its field of study, and works in a pre-interpreted world in the sense that the meanings that subjects hold are part of their construction of the world (Giddens, 1976).

The difficulty in which positivism finds itself is that it regards human behaviour as passive, essentially determined and controlled, thereby ignoring intention, individualism and freedom, i.e. as suffering from the same difficulties that inhere in behaviourism (see Chomsky's (1959) withering criticism). This problem with positivism also rehearses the familiar problem in social theory, namely, the tension between agency and structure (Layder, 1994): humans exercise agency – individual choice and intention – not necessarily in circumstances of their own choosing, but nevertheless they do not behave simply or deterministically like puppets.

Finally, the findings of positivistic social science are often said to be so banal and trivial that they are of little consequence to those for whom they are intended, namely, teachers, social workers, counsellors, managers and the like. The more effort, it seems, that researchers put into their scientific experimentation in the laboratory by restricting, simplifying and controlling variables, the more likely they are to end up with a stripped down, artificial, deterministic view of the world as if it were a laboratory.[3]

These are formidable criticisms; but what alternatives are proposed by the detractors of positivistic social science?

1.10 Post-positivism

The positivist view of the world is of an ordered, controllable, predictable, standardized, mechanistic, deterministic, stable, objective, rational, impersonal, largely inflexible, closed system whose study yields immutable, absolute, universal laws and patterns of behaviour (a 'grand narrative', a 'metanarrative') and which can be studied straightforwardly through the empirical, observational means of the scientific method. It suggests that there are laws of cause and effect, often of a linear nature (a specific cause produces a predictable effect, a small cause (stimulus) produces a small effect (response) and a large cause produces a large effect), which can be understood typically through the application of the scientific method as set out earlier in this chapter. Like a piece of clockwork, there is a place for everything and everything is in its place. It argues for an external and largely singular view of an objective reality (i.e. external to, and independent of, the researcher) that is susceptible to scientific discovery and laws. However, as Lukenchuk (2013) notes, positivism has been discarded as a useful scientific paradigm as it has failed to provide a 'logically unified system of theoretical statements grounded in the certainty of sense experience' (p. 16) and has been superseded by post-positivism.

Post-positivists challenge the positivist view of the world. Here, following Popper (1968, 1980), our knowledge of the world is not absolute but partial, conjectural, falsifiable, challengeable, provisional, probabilistic and changing. Whilst still embracing the scientific method and the acceptance of an objective world, it recognizes that there is no absolute truth, or, at least, not one which is discoverable by humans, but, rather, probabilistic knowledge only. Secure, once-and-for-all foundational knowledge and grand narratives of a singular objective reality, discoverable through empiricism, positivism, behaviourism and rationalism, are replaced by tentative speculation in which multiple perspectives, claims and warrants are brought forward by the researcher (Phillips and Burbules, 2000). The world is multilayered, able to tolerate multiple interpretations, and in which – depending on the particular view of post-positivism that is being embraced – there exist multiple external realities; knowledge is regarded as subjective rather than objective. In short, the values, biographies, perceptions, theories, environment and existing knowledge of researchers influence what is being observed, and this undermines the foundationalism of empiricism with its claims to neutral sensory experience and observation (Phillips and Burbules, 2000, p. 17). As mentioned earlier, theory is underdetermined by evidence, as the same evidence can support several different theories.

Post-positivists argue that facts and observations are theory-laden and value-laden (Feyerabend, 1975; Popper, 1980; Reichardt and Rallis, 1994), facts and theories are fallible, different theories may support specific observations/facts, and social facts, even ways of thinking and observing, are social constructions rather than objectively and universally true (Nisbett, 2005).

Imagine that a researcher observes a class lesson and notices one student winking at the teacher. Is this student being cheeky (a theory of deviant or challenging behaviour), a sign of understanding (a theory of cognition/recognition), a physical problem (Tourette's syndrome), a sign of stress or happiness (a theory of emotional behaviour), a sign of friendliness (a theory of interpersonal non-verbal behaviour), or what? The

observation on its own cannot tell us. There is a gap between an observed phenomenon and the explanation or theory of, or a hypothesis about, the phenomenon. As Phillips and Burbules (2000, pp. 18–19) remark, phenomena do not speak for themselves. This gap cannot be bridged by observed evidence alone, but needs help from outside that observed phenomenon, i.e. from non-sensory experience. What we see depends on our viewpoint. This is not to say that there is no correct answer or that multiple interpretations are acceptable (relativism), only that the observation alone is not sufficient to denote meaning.

Out goes foundational knowledge and in comes non-foundational, tentative, conjectural speculation and probabilistic, fallibilistic, imperfect, context-bound knowledge of multiple truths of a situation and multiple realities, whose validity has to be warranted whilst recognizing that such warrants may be overturned in light of future evidence. Here the separation of fact and value in positivism is unsustainable, and the foundationalism of empiricism is replaced by an admission that observation is theory-laden, and our values, perspectives, paradigms, conceptual schemes, even research communities determine what we focus on, how we research, what we deem to be important, what counts as knowledge, what research 'shows', how we interpret research findings and what constitutes 'good' research.

Post-positivism argues for the continuing existence of an objective reality, i.e. it rejects relativism, but it adopts a pluralist view of multiple, coexisting realities rather than a single reality. Imagine that two people are observing a classroom; one sits at the back of the room, and the other at the front. What they see may differ, but it is still the same classroom. Multiple views are not the same as relativism; multiple truths can coexist. There is an objective reality: the classroom, but there are different views of this, i.e. 'truth' is not simply what one of the observers takes it to be, and one frame of reference may differ from another. This raises the issue of bias and value-neutrality in educational research, which we discuss in Chapter 3.

Post-positivism recognizes that we know the world only probabilistically and imperfectly. Whilst not rejecting the value of the scientific method (e.g. experimentation), it argues for the reformulation of the strength of theories and claims made from the scientific method, namely, that their strengths are contingent on their ability to withstand 'severe tests' of their falsifiability and that their discoveries are subject to future falsification in the light of new evidence. Seen in this light, the gap between natural sciences and social science evaporates. In the post-positivist view of science, characterized by the theory-laden nature of observations, the underdetermination of theory by empirical evidence, the importance of the community of scholars in validating warrants for knowledge, the tentative, conjectural nature of conclusions, and the multiple nature of reality and 'truths', the researcher in the natural sciences is in no more or less a privileged position than the social science researcher.

1.11 Alternatives to positivistic and post-positivist social science: naturalistic and interpretive approaches

Although opponents of positivism within social science subscribe to a variety of schools of thought, each with its own different epistemological viewpoint, they are united by their common rejection of the belief that human behaviour is governed by general, universal laws and characterized by underlying regularities. Moreover, they would agree that the social world can only be understood from the standpoint of the individuals who are part of the ongoing action being investigated and that their model of a person is an autonomous one, not the version favoured by positivist researchers. Such a view is allied to constructivism (Creswell, 2013) and to interpretive approaches to social science (discussed below).

In rejecting the viewpoint of the detached, objective observer – a mandatory feature of traditional research – anti-positivists and post-positivists would argue that individuals' behaviour can only be understood by the researcher sharing their frame of reference: understanding of individuals' interpretations of the world around them has to come from the inside, not the outside. Social science is thus seen as a subjective rather than an objective undertaking, as a means of dealing with the direct experience of people in specific contexts, where social scientists understand, explain and demystify social reality through the eyes of different participants; the participants themselves define the social reality (Beck, 1979). This is not to say that understanding subjective meanings is the only route for the researcher. Rather it is both a question of emphasis and a recognition that there are external matters that impinge on subjective meaning-making and, indeed, that what constitutes 'subjectivity' is open to question and to multiple interpretations and consequences, rather than being a unified, coherent singularity (Kettley, 2012, pp. 78–9). Subjective meanings may be as empirically testable as objective statements.

The anti-positivist/post-positivist movement has many hues, for example, postmodernism, post-structuralism

and Wittgenstein's work on language games. These have influenced areas of social science such as psychology, social psychology and sociology. In the case of psychology, for instance, a school of humanistic psychology has emerged alongside the coexisting behaviouristic and psychoanalytic schools. Arising as a response to the challenge to combat the growing feelings of dehumanization which characterize many social and cultural milieux, it sets out to study and understand the person as a *whole* (Buhler and Allen, 1972). Humanistic psychologists present a model of people that is positive, active and purposive, and at the same time stresses their own involvement with the life experience itself. They do not stand apart, introspective, hypothesizing. Their interest is directed at the intentional and creative aspects of the human being. The perspective adopted by humanistic psychologists is naturally reflected in their methodology. They are dedicated to studying the individual in preference to the group, and consequently prefer idiographic approaches to nomothetic ones. The implications of the movement's philosophy for education have been drawn by Carl Rogers (1942, 1945, 1969).

Comparable developments within social psychology may be perceived in the 'science of persons' movement. It is argued here that we must use ourselves as a key to our understanding of others and, conversely, our understanding of others as a way of finding out about ourselves, an anthropomorphic model of people. Since anthropomorphism means, literally, the attribution of human form and personality, the implied criticism is that social psychology as traditionally conceived has singularly failed, so far, to model people as they really are, and that social science should treat people as capable of monitoring and arranging their own actions, exercising their agency (Harré and Secord, 1972).

Social psychology's task is to understand people in the light of this anthropomorphic model. Proponents of this 'science of persons' approach place great store on the systematic and painstaking analysis of social episodes, i.e. behaviour in context. In Box 1.5 we give an example of such an episode taken from a classroom study. Note how the particular incident would appear on an interaction analysis coding sheet of a researcher employing a positivistic approach. Note, too, how this slice of classroom life can only be understood by knowledge of the specific organizational background and context in which it is embedded.

BOX 1.5 A CLASSROOM EPISODE

Walker and Adelman describe an incident in the following manner:

> In one lesson the teacher was listening to the boys read through short essays that they had written for homework on the subject of 'Prisons'. After one boy, Wilson, had finished reading out his rather obviously skimped piece of work the teacher sighed and said, rather crossly:
>
> T: Wilson, we'll have to put you away if you don't change your ways, and do your homework. Is that all you've done?
> P: Strawberries, strawberries. (Laughter)

Now at first glance this is meaningless. An observer coding with Flanders Interaction Analysis Categories (FIAC) would write down:

'7' (teacher criticizes) followed by a,
'4' (teacher asks question) followed by a,
'9' (pupil irritation) and finally a,
'10' (silence or confusion) to describe the laughter.

Such a string of codings, however reliable and valid, would not help anyone to *understand* why such an interruption was funny. Human curiosity makes us want to know *why* everyone laughs – and so, I would argue, the social scientist needs to know too. Walker and Adelman asked subsequently why 'strawberries' was a stimulus to laughter and were told that the teacher frequently said the pupils' work was 'like strawberries – good as far as it goes, but it doesn't last nearly long enough'. Here a casual comment made in the past has become an integral part of the shared meaning system of the class. It can only be comprehended by seeing the relationship as developing over time.

Source: Adapted from Delamont (1976)

The approach to analysing social episodes in terms of the 'actors' themselves is known as the 'ethogenic method'.[4] Unlike positivistic social psychology which ignores or presumes its subjects' interpretations of situations, ethogenic social psychology concentrates on the ways in which persons construe their social world. By probing their accounts of their actions, it endeavours to come up with an understanding of what those persons were doing in the particular episode.

As an alternative to positivist approaches, naturalistic, qualitative, interpretive approaches of various hue possess particular distinguishing features:

- people are deliberate and creative in their actions, they act intentionally and make meanings in and through their activities (Blumer, 1969);
- people actively construct their social world – they are not the 'cultural dopes' or passive dolls of positivism (Becker, 1970; Garfinkel, 1967);
- situations are fluid and changing rather than fixed and static; events and behaviour evolve over time and are richly affected by context – they are 'situated activities';
- events and individuals are unique and largely non-generalizable;
- a view that the social world should be studied in its natural state, without the intervention of, or manipulation by, the researcher (Hammersley and Atkinson, 1983);
- fidelity to the phenomena being studied is fundamental;
- people interpret events, contexts and situations, and act on the bases of those events (echoing Thomas's (1928) famous dictum that if people define their situations as real then they are real in their consequences – if I believe there is a mouse under the table, I will act as though there is a mouse under the table, whether there is or not (Morrison, 1998));
- there are multiple interpretations of, and perspectives on, single events and situations;
- reality is multilayered and complex;
- many events are not reducible to simplistic interpretation, hence 'thick descriptions' (Geertz, 1973) are essential rather than reductionism; that is to say, thick descriptions representing the complexity of situations are preferable to simplistic ones;
- researchers need to examine situations through the eyes of participants rather than the researcher.

The anti-positivist/post-positivist movement in sociology is represented by three schools of thought – phenomenology, ethnomethodology and symbolic interactionism. A common thread running through the three schools is a concern with phenomena, that is, the things we directly apprehend through our senses as we go about our daily lives, together with a consequent emphasis on qualitative as opposed to quantitative methodology. The differences between them and the significant roles each phenomenon plays in educational research are such as to warrant a more extended consideration of them in the discussion below.

1.12 A question of terminology: the normative and interpretive paradigms

So far we have introduced and used a variety of terms to describe the numerous branches and schools of thought embraced by the positivist and anti-positivist viewpoints. As a matter of convenience and as an aid to communication, we clarify at this point two generic terms conventionally used to describe these two perspectives and the categories subsumed under each, particularly as they refer to social psychology and sociology. The terms in question are 'normative' and 'interpretive'. The normative paradigm (or model) contains two major orienting ideas (Douglas, 1973): first, that human behaviour is essentially rule-governed; and second, that it should be investigated by the methods of natural science. The interpretive paradigm, in contrast to its normative counterpart, is characterized by a concern for the individual. Whereas normative studies are positivist, theories constructed within the context of the interpretive paradigm tend to be anti-positivist. As we have seen, the central endeavour in the context of the interpretive paradigm is to understand the subjective world of human experience. To retain the integrity of the phenomena being investigated, efforts are made to get inside the person and to understand from within. The imposition of external form and structure is resisted, since this reflects the viewpoint of the observer as opposed to that of the actor directly involved.

Two further differences between the two paradigms may be identified here: the first concerns the concepts of 'behaviour' and 'action'; the second, the different conceptions of 'theory'. A key concept within the normative paradigm, 'behaviour' refers to responses either to external environmental stimuli (e.g. another person, or the demands of society) or to internal stimuli (e.g. hunger, or the need to achieve). In either case, the cause of the behaviour lies in the past. Interpretive approaches, on the other hand, focus on action. This may be thought of as behaviour-with-meaning; it is intentional behaviour, and as such, future oriented. Actions are only meaningful to us insofar as we are able to ascertain the intentions of actors to share their

experiences. A large number of our everyday interactions with one another rely on such shared experiences.

As regards theory (see also Chapter 4), normative researchers try to devise general theories of human behaviour and to validate them through the use of research methodologies which, some believe, push them further and further from the experience and understanding of the everyday world and into a world of abstraction. For them, the basic reality is the collectivity; it is external to the actor and manifest in society, its institutions and its organizations. The role of theory is to say how reality hangs together in these forms or how it might be changed so as to be more effective. The researcher's ultimate aim is to establish a comprehensive 'rational edifice', a universal theory, to account for human and social behaviour.

But what of the interpretive researchers? They begin with individuals and set out to understand their interpretations of the world around them. Indeed they use approaches such as '*verstehen*' ('understanding') and hermeneutic (uncovering and interpreting meanings) to try to see the social world through the eyes of the participants, rather than as an outsider. Here is a view which states that, unlike natural scientists, social scientists recognize that human behaviour is intentional, that people interpret situations through their own eyes and act on those interpretations and that the research has to take cognizance of this. People make sense of the world in their own terms, and such interpretation takes place in socio-cultural, socio-temporal and socio-spatial contexts (cf. Marshall and Rossman, 2016). In turn this requires researchers to suspend or forgo their own assumptions about people, cultures and contexts in favour of looking at a situation and its context in its own terms (cf. Hammersley, 2013, p. 27), to set aside the search for universal statements or causal laws, i.e. to adopt idiographic rather than the nomothetic research of the positivists. The nature of research, then, is exploratory in nature, to investigate the interpretations of the situation made by the participants themselves, to understand their attitudes, behaviours and interactions.

In interpretive research, theory is emergent and arises from particular situations; it is 'grounded' in data generated by the research (Glaser and Strauss, 1967) (see Chapter 37). Theory should not precede research but follow it. Investigators work directly with experience and understanding to build their theory on them. The data thus yielded will include the meanings and purposes of those people who are their source. Further, the theory so generated must make sense to those to whom it applies. The aim of scientific investigation for the interpretive researcher is to understand how this reality goes on at one time and in one place and compare it with what goes on in different times and places. Thus theory becomes sets of meanings which yield insight and understanding of people's behaviour. These theories are likely to be as diverse as the meanings and understandings that they seek to explain. From an interpretive perspective, the hope of a universal theory which characterizes the normative outlook gives way to multifaceted images of human behaviour as varied as the situations and contexts supporting them.

1.13 Phenomenology, ethnomethodology, symbolic interactionism and constructionism

There are many variants of qualitative, naturalistic, interpretive approaches (Hitchcock and Hughes, 1995). Marshall and Rossman (2016) identify several such 'genres' (pp. 17–41). Under 'major genres' they include those which: (a) focus on culture and society (e.g. ethnographic approaches); (b) focus on the lived experiences of individuals (phenomenological approaches); (c) focus on texts and talking (sociolinguistic approaches); (d) use grounded theory approaches; and (e) use case studies. Under 'critical genres' they include: (a) critical ethnography and autoethnography; (b) critical discourse analysis; (c) action research and participatory action research; (d) queer theory; (e) critical race theory; (f) feminist theory; (g) cultural studies; and (h) internet/virtual ethnography. We discuss critical theories in Chapter 3. Here we focus on four significant 'traditions' in the interpretive style of research – phenomenology, ethnomethodology, symbolic interactionism and constructionism.

Phenomenology

In its broadest meaning, phenomenology is a theoretical point of view that advocates the study of direct experience taken at face value and which sees behaviour as determined by the phenomena of experience rather than by external, objective and physically described reality (English and English, 1958). Although phenomenologists differ among themselves on particular issues, there is fairly general agreement on the following points identified by Curtis (1978), Hammersley (2013) and Marshall and Rossman (2016), which can be taken as distinguishing features of their philosophical viewpoint:

- A belief in the importance, and even the primacy, of subjective consciousness;
- The importance of documenting and describing immediate experiences;

- The significance of understanding how and why participants' knowledge of a situation comes to be what it is;
- The social and cultural situatedness of actions and interactions, together with participants' interpretations of a situation;
- An understanding of consciousness as active, as meaning bestowing;
- A claim that there are certain essential structures to consciousness of which we gain direct knowledge by a certain kind of reflection. Exactly what these structures are is a point about which phenomenologists differ.

Various strands of development may be traced in the phenomenological movement: we briefly examine two of them – the transcendental phenomenology of Husserl; and existential phenomenology, of which Schutz is perhaps the most characteristic representative.

Husserl, regarded by many as the founder of phenomenology, was concerned with investigating the source of the foundation of science and with questioning the common-sense, 'taken-for-granted' assumptions of everyday life (see Burrell and Morgan, 1979). To do this, he set about opening up a new direction in the analysis of consciousness. His catchphrase was 'back to the things!' which for him meant finding out how things appear directly to us rather than through the media of cultural and symbolic structures. In other words, we are asked to look beyond the details of everyday life to the essences underlying them. To do this, Husserl exhorts us to 'put the world in brackets' or free ourselves from our usual ways of perceiving the world. What is left over from this reduction is our consciousness, of which there are three elements – the 'I' who thinks, the mental acts of this thinking subject, and the intentional objects of these mental acts. His was a call to overcome the subjective–objective divide. The aim, then, of this method of *epoché*, as Husserl called it, is the dismembering of the constitution of objects in such a way as to free us from all preconceptions about the world.

Schutz was concerned with relating Husserl's ideas to the issues of sociology and to the scientific study of social behaviour. Of central concern to him was the problem of understanding the meaning structure of the world of everyday life. He sought the origins of meaning in the 'stream of consciousness' – basically an unbroken stream of lived experiences which have no meaning in themselves. One can only impute meaning to them retrospectively, by the process of turning back on oneself and looking at what has been going on. In other words, meaning can be accounted for here by the concept of reflexivity. For Schutz, the attribution of meaning reflexively is dependent on the people identifying the purpose or goal they seek (Burrell and Morgan, 1979).

According to Schutz, the way we understand the behaviour of others is dependent on a process of typification by means of which the observer makes use of concepts resembling 'ideal types' to make sense of what people do. These concepts are derived from our experience of everyday life and it is through them, claims Schutz, that we classify and organize our everyday world. In this respect he adhered to principles of empiricism. As Burrell and Morgan observe, we learn these typifications through our biographical locations and social contexts. Our knowledge of the everyday world inheres in social order and itself is socially ordered.

The fund of everyday knowledge by means of which we are able to typify other people's behaviour and come to terms with social reality varies from situation to situation. We thus live in a world of multiple realities, and social actors move within and between these, abiding by the rules of the game for each of these worlds.

Ethnomethodology

Like phenomenology, ethnomethodology is concerned with the world of everyday life, studying participants' circumstances, thoughts and commonplace daily lives as worthy of empirical study (Garfinkel, 1967, p. vii). Garfinkel maintains that students of the social world must doubt the reality of that world; and that in failing to view human behaviour more sceptically, sociologists have created an ordered social reality that bears little relationship to the real thing. He thereby challenges the basic sociological concept of order.

Ethnomethodology, then, is concerned with how people make sense of their everyday world. More especially, it is directed at the mechanisms by which participants achieve and sustain interaction in a social encounter – the assumptions they make, the conventions they utilize, and the practices they adopt. Ethnomethodology thus seeks to understand social accomplishments in their own terms; it is concerned to understand them from within (Burrell and Morgan, 1979).

In identifying the 'taken-for-granted' assumptions characterizing a social situation and the ways in which the people involved make their activities rationally accountable, ethnomethodologists use notions of 'indexicality' and 'reflexivity'. Indexicality refers to the ways in which actions and statements are related to the social contexts producing them, and to the way

their meanings are shared by the participants but not necessarily stated explicitly. Indexical expressions are thus the designations imputed to a particular social occasion by the participants in order to locate the event in the sphere of reality. Reflexivity, on the other hand, refers to the way in which all accounts of social settings – descriptions, analyses, criticisms etc. – and the social settings occasioning them, are mutually interdependent.

One can distinguish between two types of ethnomethodologists: *linguistic* and *situational*. *Linguistic* ethnomethodologists focus upon the use of language and the ways in which conversations in everyday life are structured. Their analyses make much use of the unstated 'taken-for-granted' meanings, the use of indexical expressions and the way in which conversations convey much more than is actually said. *Situational* ethnomethodologists cast their view over a wider range of social activity and seek to understand the ways in which people negotiate the social contexts in which they find themselves. They are concerned to understand how people make sense of and order their environment. As part of their empirical method, ethnomethodologists may consciously and deliberately disrupt or question the ordered 'taken-for-granted' elements in everyday situations in order to reveal the underlying processes at work.

The substance of ethnomethodology thus largely comprises a set of specific techniques and approaches to be used in studying what Garfinkel has described as the 'awesome indexicality' of everyday life. It is geared to empirical study, and the stress which its practitioners place upon the uniqueness of the situation encountered projects its essentially relativist standpoint. A commitment to the development of methodology and fieldwork has occupied first place in the interests of its adherents, so that related issues of ontology, epistemology and the nature of human beings have received less attention than perhaps they deserve.

Symbolic interactionism

Essentially, the notion of symbolic interactionism derives from the work of Mead (1934). Although subsequently to be associated with such noted researchers as Blumer, Hughes, Becker and Goffman, the term does not represent a unified perspective in that it does not embrace a common set of assumptions and concepts accepted by all who subscribe to the approach. Here, however, it is possible to identify three basic postulates. These have been set out by Woods (1979) as follows. First, human beings act towards things on the basis of the meanings they have for them. Humans inhabit two different worlds: the 'natural' world wherein they are organisms of drives and instincts and where the external world exists independently of them, and the social world where the existence of symbols, like language, enables them to give meaning to objects. This attribution of meanings, this interpreting, is what makes them distinctively human and social. Interactionists therefore focus on the world of subjective meanings and the symbols by which they are produced and represented. This means not making any prior assumptions about what is going on in an institution, and taking seriously, indeed giving priority to, inmates' own accounts. Thus, if students appear preoccupied for too much of the time – 'being bored', 'having a laugh' etc. – the interactionist is keen to explore the properties and dimensions of these processes.

Second, this attribution of meaning to objects through symbols is a continuous process. Action is not simply a consequence of psychological attributes such as drives, attitudes or personalities, or determined by external social facts such as social structure or roles, but results from a continuous process of meaning attribution which is always emerging, in a state of flux and subject to change. The individual constructs, modifies, pieces together, weighs up the pros and cons, and bargains.

Third, this process takes place in a social context. Individuals align their actions to those of others. They do this by 'taking the role of the other', by making indications to themselves about others' likely responses. They construct how others wish to or might act in certain circumstances, and how they themselves might act. They might try to 'manage' the impressions others have of them, put on a 'performance', try to influence others' 'definition of the situation'.

Instead of focusing on the individual, then, and his or her personality characteristics, or on how the social structure or social situation causes individual behaviour, symbolic interactionists direct their attention at the nature of interaction, the dynamic activities taking place between people. In focusing on the interaction itself as a unit of study, the symbolic interactionist creates a more active image of the human being and rejects the image of the passive, determined organism. Individuals interact; societies are made up of interacting individuals. People are constantly undergoing change in interaction and society is changing through interaction. Interaction implies human beings acting in relation to each other, taking each other into account, acting, perceiving, interpreting, acting again. Hence, a more dynamic and active human being emerges rather than an actor merely responding to others. Woods (1983, pp. 15–16) summarizes key emphases of symbolic interaction thus:

- individuals as constructors of their own actions;
- the various components of the self and how they interact; the indications made to self, meanings attributed, interpretive mechanisms, definitions of the situation; in short, the world of subjective meanings, and the symbols by which they are produced and represented;
- the process of negotiation, by which meanings are continually being constructed;
- the social context in which they occur and whence they derive;
- by taking the 'role of the other' – a dynamic concept involving the construction of how others wish to or might act in a certain circumstance, and how individuals themselves might act – individuals align their actions to those of others.

Constructionism

In constructionism (also termed constructivism), in contrast to the argument that external objects and factors determine, shape, impress, print or fix themselves onto passive recipients (i.e. are 'givens' in society or individuals), people actively and agentically seek out, select and construct their own views, worlds and learning, and these processes are rooted in socio-cultural contexts and interactions. In other words, cognition is generative and active rather than receptive and passive respectively. Through such active cognition and deliberate perception we come to understand ourselves and how this affects the worlds we inhabit and the way in which we interact with the objects and people in them.

Hammersley (2013) notes that constructionism requires researchers to focus on the processes that lead to the construction, constitution and character given to independent objects and the relationships between them (pp. 35–6), i.e. how people collectively construct their social worlds (e.g. through discourse analysis) (p. 36). He gives an example of replacing the definition of a person as 'intelligent' with an examination of the 'discursive practices' which led to the construction of that person being intelligent and how this affects how that person operates in socio-cultural and institutional contexts (p. 36).

Social constructionism holds that individuals seek to make meaning of their social lives and that the researcher has to examine the situation in question through the multiple lenses of the individuals involved, to obtain their definition of the situation, to see how they make sense of their situation and to focus on interactions, contexts, environments and biographies. Indeed social constructionism emphasizes the social nature of learning, arguing that it is only through social interaction and communication that certain types of learning occur and certain views of the world are constructed.

A characteristic common to the phenomenological, ethnomethodological, symbolic interactionist and constructionist perspectives, which makes them attractive to the educational researcher, is the way they fit naturally to the kind of concentrated action found in classrooms and schools. Yet another shared characteristic is the manner in which they are able to preserve the integrity of the situation in which they are employed. Here the influence of the researcher in structuring, analysing and interpreting the situation is present to a much smaller degree than would be the case with a more traditionally oriented research approach.

1.14 Criticisms of the naturalistic and interpretive approaches

Critics have wasted little time in pointing out what they regard as weaknesses in these newer qualitative perspectives. They argue that while it is undeniable that our understanding of the actions of our fellow-beings necessarily requires knowledge of their intentions, this, surely, cannot be said to constitute *the* purpose of a social science. As Rex observed:

> Whilst patterns of social reactions and institutions may be the product of the actors' definitions of the situations there is also the possibility that those actors might be falsely conscious and that sociologists have an obligation to seek an objective perspective which is not necessarily that of any of the participating actors at all.… We need not be confined purely and simply to that … social reality which is made available to us by participant actors themselves.
>
> (Rex, 1974)

While these more recent perspectives have presented models of people that are more in keeping with common experience, some argue that anti-positivists/post-positivists have gone too far in abandoning scientific procedures of verification and in giving up hope of discovering useful generalizations about behaviour. Are there not dangers in rejecting the approach of physics in favour of methods more akin to literature, biography and journalism? Some specific criticisms of the methodologies are well directed, for example Argyle (1978) questions whether, if carefully controlled interviews such as those used in social surveys are inaccurate, then the less controlled interviews carry even greater risks of inaccuracy. Indeed Bernstein (1974) suggests that subjective reports may be incomplete and misleading.

I may believe that the teacher does not like me, and, therefore, act as though the teacher does not like me (a self-fulfilling prophecy), but, in fact, all the time the teacher actually does like me; my perception is wrong.

Bernstein's criticism is directed at the overriding concern of phenomenologists and ethnomethodologists with the meanings of situations and the ways in which these meanings are negotiated by the actors involved. What is overlooked about such negotiated meanings, observes Bernstein, is that the very process whereby one interprets and defines a situation is itself a product of the circumstances in which one is placed. One important factor in such circumstances that must be considered is the power of others to impose their own definitions of situations upon participants. Doctors' consulting rooms and headteachers' studies are locations in which inequalities in power are regularly imposed upon unequal participants. The ability of certain individuals, groups, classes and authorities to persuade others to accept their definitions of situations demonstrates that while – as ethnomethodologists insist – social structure is a consequence of the ways in which we perceive social relations, it is clearly more than this.

Conceiving of social structure as external to ourselves helps us include its self-evident effects upon our daily lives into our understanding of the social behaviour going on about us. Here is rehearsed the tension between agency and structure of social theorists (Layder, 1994); the danger of interactionist and interpretive approaches is their relative neglect of the power of external – structural – forces to shape behaviour and events. There is a risk in interpretive approaches that they become hermetically sealed from the world outside the participants' theatre of activity – they put artificial boundaries around subjects' behaviour. Just as positivistic theories can be criticized for their macro-sociological persuasion, so interpretive and qualitative theories can be criticized for their narrowly micro-sociological perspectives.

1.15 Postmodernist and post-structuralist perspectives

It is not only post-positivists who challenge the modernist, positivist conception of the world. For modernists the world is available to be studied objectively and, by using scientific methods, to arrive at secure, rigorous, scientific, discipline-based explanations of observed phenomena – 'grand narratives' which are redolent of the Enlightenment project of providing foundationalist and absolute knowledge. Postmodernism challenges each of these. Whilst it is perhaps invidious to try to characterize postmodernists (as they would argue against any singular or all-embracing definitions), in a seminal text Jameson (1991) argues that postmodernism does have several distinguishing hallmarks, including, for example:

- the absence of 'grand narratives' (metanarratives) and grand designs, laws and patterns of behaviour (thereby, ironically, eclipsing the status of their own narrative);
- the valorization of discontinuity, difference, diversity, pluralism, variety, uniqueness, subjectivity, distinctiveness and individuality;
- the importance of the local, the individual and the particular;
- the 'utter forgetfulness of the past' and the 'autoreferentiality' of the present (Jameson, 1991, p. 42);
- the importance of temporality and context in understanding phenomena: meanings are rooted in time, space, cultures, societies and are not universal across these;
- the celebration of depthlessness, multiple realities (and, as Jameson argues, multiple superficialities) and the rectitude of individual interpretations and meanings rather than an appeal to a singular or universal rationalism;
- relativism rather than absolutism in deciding what constitutes worthwhile knowledge, research and their findings;
- the view of knowledge as a human, social construct;
- multiple, sometimes contradictory, yet co-existent interpretations of the world, in which the researcher's interpretation is only one out of several possible interpretations, i.e. the equal value of different interpretations and the reduction in the authority of the researcher, yet, simultaneously, the privileging of some interpretations of the world to the neglect of others (i.e. the nexus between knowledge and power, a feature of critical theory, discussed in Chapter 3);
- the recognition that researchers are part of the world that they are researching;
- the emancipatory potential of according value to individual views, values, perspectives and interpretations (see Chapter 3).

Pring (2015) adds to this the point that postmodernism is characterized by a revolt against thought control and cultural control, by an assertion of multiple forms of cultural expression, an abandonment of certainty, a replacement of 'authority' (as in 'authoritative') by multiple voices and negotiated meanings, and a blurring of artificial boundaries (disciplines) of knowledge, a questioning of received wisdoms and a recognition of

fallibilism, all of which he sees as the function of the 'perennial philosophical tradition' and not one given birth to by postmodernism (pp. 134–7). In one sense postmodernism supports the interpretive paradigm set out earlier in this chapter. In another sense it supports complexity theory as discussed below, and in a third sense it supports critical theory as set out in Chapter 3. Postmodernism has a chameleon-like nature in this respect.

Post-structuralism, like postmodernism, has many different interpretations (we will not discuss here the interpretation that relates to semiology). Here we take a necessarily selective interpretation, to focus on those features that are relevant to the foundations and conduct of educational research. Here post-structuralism can be regarded as a counter to those structural-functionalists who adopt a systems view of society (e.g. Marxism, or functionalist anthropologists such as Lévi-Strauss) or behaviour as a set of interrelated parts which, in law-like fashion, pattern themselves and fit together neatly into a fixed view of the world and its operations and in which individual behaviour is largely determined by given, structural features of society (e.g. social class, position in society, role in society). In post-structuralist approaches, data (e.g. conversations, observations) and even artefacts can be regarded as texts (Burman and Parker, 1993), as discourses that are constructed and performed through discourses (see Chapter 35), open to different meaning and interpretations (Francis, 2010, p. 327).

Post-structuralists (e.g. Foucault, Derrida) argue that individual agency has prominence; individuals are not simply puppets of a given system; people are diverse and different, indeed they may carry contradictions and tensions within themselves (e.g. in terms of class, ethnicity, gender, employment, social group, family membership and tasks, and so on); they are not simply the decentred bearers of given roles. Individuals have views of themselves, and one task of the researcher is to locate research findings within the views of the self that the participants hold, and to identify the meanings which the participants accord to phenomena. Hence not only do the multiple perspectives of the participants have to be discerned, but also those of the researchers, the audiences of the research and the readers of research. The task of the research is to 'deconstruct', to expose, the different meanings, layers of meanings and privileging of meanings inherent in a phenomenon or piece of research. There is no single, 'essential' meaning, but many, and one task of research is to understand how meanings and knowledge are produced, legitimized and used. (This links post-structuralism to critical theory, though some critical theorists, e.g. Habermas (1987), argue against critical theory's affinity to postmodernism or post-structuralism.)

One can detect affinities between post-positivism, postmodernism and post-structuralism in underpinning interpretive and qualitative approaches to educational research, complexity theory and critical theory, and the significance given to individual and subjective accounts in the research process, along with reflexivity on the part of the researcher. (That said, many post-positivists, postmodernists and post-structuralists would reject such a simple affinity, or even the links between their views and, for example, phenomenology and interpretivism. We do not explore this here.) One can suggest that post-positivism, postmodernism and post-structuralism argue for multiple interpretations of a phenomenon to be provided, to accord legitimacy to individual voices in research, and to abandon the search for deterministic, simple cause-and-effect laws of behaviour and action.

1.16 Subjectivity and objectivity in educational research

The preceding overview has alluded to the sympathies between some paradigms and objectivity in research and other paradigms and subjectivity in research. To make such an exclusive separation is a chimera, a false dichotomy. With regard to objectivity, to say, for instance, that objectivity inheres in positivist and post-positivist approaches overlooks not only the several interpretations of positivism and post-positivism but what it means to be subjective. Objectivity is refracted through the researcher's eyes and the generation, construction and testing of hypotheses draw on personal understandings and formulations. In other words, objectivity cannot escape some subjective roots. Taken to an extreme, it leads to a rejection of the idea that the researcher can ever be objective, just as there is a rejection of the idea that there is an objective reality or 'truth' about a phenomenon (Hammersley, 2011, p. 89). Objectivity here is defined as intersubjectivity (as opposed to subjectivity), reliability and freedom from bias (Risjord, 2014, p. 22). Risjord illustrates the difference between intersubjectivity and subjectivity thus (p. 23): I feel hungry (subjective) so I eat a sandwich (intersubjective, in that it can be seen by an observer, i.e. is open to critical scrutiny).

On the other hand, subjectivity cannot turn its back on what is 'out there' in terms of overriding the social, societal and institutional social facts, which have an existence independent of the participant. Subjectivity cannot lay claim to being a privileged discourse without risking relativism.

Subjectivity and objectivity are frequently placed at the poles of different continua (cf. Hammersley, 2011, p. 90), for example:

Subjective		*Objective*
Internal	… … … … . .	External
Private	… … … … . .	Public
Positivist	… … … … . .	Interpretive
Idiographic	… … … … . .	Nomothetic
Judgement	… … … … . .	Technical application (e.g. of statistics)
Individual	… … … … . .	Shared
Personal	… … … … . .	Impersonal
Particular	… … … … . .	General
Relative	… … … … . .	Absolute
Opinion	… … … … . .	Proof
Experimental	… … … … . .	Interactionist
Biased	… … … … . .	Bias-free
Unobservable	… … … … . .	Observable
Idiosyncratic	… … … … . .	Regular
Uncertain	… … … … . .	Certain
Unpredictable	… … … … . .	Predictable
Unreliable	… … … … . .	Reliable
Imprecise	… … … … . .	Explicit
Questionable	… … … … . .	Conclusive
Unverifiable	… … … … . .	Checkable
Prone to error	… … … … . .	Secure
Complex	… … … … . .	Straightforward
Opaque	… … … … . .	Transparent

Source: Adapted from Barr Greenfield (1975)

However, this creates false dichotomies, and look how easily one can create biases in the pejorative terms used: many of the items in the left-hand column are presented as the shabby, less respectable end of research, whilst the right-hand column seems much more clean and respectable. This can overlook the risk of bias and errors that researchers might commit in working in the right-hand column and the authenticity, correctness and truth of the left-hand column. Both subjective and objective views have to face judgements of plausibility, validity, reliability, meaningfulness and credibility.

However, more fundamentally, as Hammersley (2011) remarks, we depend on personal knowledge and judgement in making meaning of phenomena and data, be those data numbers, words, pictures or sounds. We rely on our senses in making observations. Following objective procedures requires a personal commitment. We rely on our judgement in raising hypotheses, making inferences and drawing conclusions. However, simply amassing subjective data from participants does not ensure that the data are true or reliable, but stating objective procedures does not ensure identical practices, not least as, in the social world, researchers – consciously or not – adjust their practices to the situation and the people who are participating in the research; standardizing practice has to extend to participants.

Medical research is a good example here: whilst there might be an objective, standardized procedure for patients taking medicine in a randomized controlled trial, that does not guarantee that patients will follow it: they might refuse to take the medicine, forget to take it, take it at the wrong time of day, take some but not all of it, take the wrong dose (too little or too much), misread the instructions, and so on. Intention does not match actuality.

The claims we make from knowledge, be they from the left-hand or right-hand columns here, do not constitute absolute truth: the same data can, and do, sustain multiple interpretations, claims and conclusions. Further, is it really possible or desirable to set aside one's own biography, values and assumptions, however reflexive one might be? Reflexivity is not the same as objectivity. Is it not the case, anyway, that knowledge, particularly of the social world, is a socio-temporal construction rather than the clean world of the objectivist, and to pretend otherwise is simply naive or deceitful (Hammersley, 2011, p. 96)? Or is this giving in to the relativists and the postmodernists, in the knowledge that relativism is, by its own definition, only relative, and that the postmodernists cannot lay claim to their views as having any status at all as to do so would be to acknowledge that metanarratives exist – a claim which postmodernists proscribe as an article of faith.

Hammersley (2011) is clear that errors may stem from the researcher's own social or individual characteristics and their influence on their research, but that it is unnecessary and, indeed, undesirable to assume that the researcher can stand out completely from his or her social and individual characteristics. Further, error does not automatically follow from an acknowledgement of the researcher's own social and individual characteristics.

The task, then, is to protect the research from negative effects of subjectivity (2011, p. 101), though Hammersley acknowledges that what constitutes 'error' is not always clear. However, he offers researchers some advice here, cautioning them to be on their guard against preconceptions, prior assumptions, preferences and biases that are 'external to the pursuit of knowledge' (p. 102), i.e. which are goals that are separate

from the research itself. Objectivity, in this case, means adhering to the 'epistemic virtue' of keeping *only* to the canons and requirements of the research itself, setting aside any extraneous personal convictions or subordinating the research to any other goals outside the research (p. 103). Given this, objectivity and the suppression of personal, subjective beliefs, values, commitments or agendas have a key role to play in educational research. The objective reliability of the research does not depend on the political, valuative or moral motivations of the researcher (cf. Risjord, 2014, p. 23).

Similarly, value-neutrality in educational and social science research leaves unsaid any comment on what *ought* or *ought not* be done; that is for policy makers. Rather, educational research confines itself to facts; that is, the scientific enterprise. Saying that teachers should not assault students is an evaluative statement and not a matter for social science research, as it does not rest on empirical data alone, though reporting incidents of assault and its effects surely is a matter for research.

Whether researchers should have a 'committed' position is a matter that we return to in Chapter 3 on critical theory, which explicitly disavows value-free positions, and argues for partisan positions in research as contributing to the greater good of an emancipated society in freeing itself from that ideology which conceals oppression and unjust subordination and power differentials of social groups, and which transforms society to equality, democracy and social justice. Fact and value reunite.

1.17 The paradigm of complexity theory

An emerging paradigm in educational research is that of complexity theory (Medd, 2002; Morrison, 2002a, 2008; Radford, 2006, 2007, 2008; Kuhn, 2007; Byrne and Callaghan, 2014; Boulton *et al*., 2015), as schools can be regarded as 'complex adaptive systems' (Kaufmann, 1995). Complexity theory looks at the world in ways which break with simple cause-and-effect models, simple determinism and linear predictability (Morrison, 2008) and a dissection/atomistic approach to understanding phenomena (Radford, 2007, 2008; Byrne and Callaghan, 2014), replacing them with organic, non-linear and holistic approaches (Santonus, 1998, p. 3). Relations within interconnected, dynamic and changing networks are the order of the day (Wheatley, 1999, p. 10), and there is a 'multiplicity of simultaneously interacting variables' (Radford, 2008, p. 510). Here key terms are feedback, recursion, emergence, connectedness and self-organization. Out go the simplistic views of linear causality (Radford, 2007; Morrison, 2009; Byrne and Callaghan, 2014; Boulton *et al*., 2015), the ability to predict, control and manipulate, to apply reductive techniques to research, and in come uncertainty, networks and connection, holism self-organization, emergence over time through feedback and the relationships of the internal and external environments, and survival and development through adaptation and change.

In complexity theory, a self-organizing system is autocatalytic and possesses its own unique characteristics and identity (Kelly and Allison, 1999, p. 28) which enable it to perpetuate and renew itself over time – it creates the conditions for its own survival. This takes place through engagement with others in a system (Byrne and Callaghan, 2014; Boulton *et al*., 2015). The system is aware of its own identity and core properties, and is self-regenerating (able to sustain that identity even though aspects of the system may change, e.g. staff turnover in a school).

Through feedback, recursion, perturbance, autocatalysis, connectedness and self-organization, higher levels of complexity and differentiated, new forms of life, behaviour and systems arise from lower levels of complexity and existing forms. These complex forms derive from often comparatively simple sets of rules – local rules and behaviours generating emergent complex global order and diversity (Waldrop, 1992, pp. 16–17; Lewin, 1993, p. 38). General laws of emergent order can govern adaptive, dynamical processes (Waldrop, 1992, p. 86; Kauffman, 1995, p. 27).

The interaction of individuals feeds into the wider environment which, in turn, influences the individual units of the network; they co-evolve, shaping each other (Stewart, 2001), and co-evolution requires connection, cooperation and competition: competition to force development and cooperation for mutual survival. The behaviour of a complex system as a whole, formed from its several elements, is greater than the sum of the parts (Byrne and Callaghan, 2014; Boulton *et al*., 2015).

Feedback occurs between the interacting elements of the system. Negative feedback is regulatory (Marion, 1999, p. 75), for example learning that one has failed a test. Positive feedback brings increasing returns and uses information to change, grow and develop (Wheatley, 1999, p. 78); it amplifies small changes (Stacey, 1992, p. 53). Once a child has begun to read she is gripped by reading, she reads more and learns at an exponential rate.

Connectedness, a key feature of complexity theory, exists everywhere. In a rainforest ants eat leaves, birds eat ants and leave droppings, which fertilize the soil for

growing trees and leaves for the ants (Lewin, 1993, p. 86). In schools, children are linked to families, teachers, peers, societies and groups; teachers are linked to other teachers, support agencies (e.g. psychological and social services), policy-making bodies, funding bodies, the legislature, and so on. The child (indeed the school) is not an island, but is connected externally and internally in several ways. Disturb one element and the species or system must adapt or die; the message is ruthless.

Emergence is the partner of *self-organization*. Systems possess the ability for self-organization, which is not according to an a priori grand design – a cosmological argument – nor a teleological argument; complexity is neither. Further, self-organization emerges, it is internally generated; it is the opposite of external control. As Kauffman (1995) suggests, order comes for free and replaces control. Order is not imposed; it emerges; in this way it differs from control. Self-organized order emerges of itself as the result of the interaction between the organism and its environment, and new structures emerge that could not have been predicted; that emerged system is, itself, complex and cannot be reduced to those parts that gave rise to the system. As Davis and Sumara (2005, p. 313) write: 'phenomena have to be studied at their level of emergence', i.e. at their present overall state, not in terms of the elements present in the pre-metamorphosed state.

Stacey (2000) suggests that a system can only evolve, and evolve spontaneously, where there is diversity and deviance (p. 399) – a salutary message for command-and-control teachers who exact compliance from their pupils. The future is largely unpredictable. At the point of 'self-organized criticality' (Bak, 1996), a tipping point, the effects of a single event are likely to be very large, breaking the linearity of Newtonian reasoning wherein small causes produce small effects; the straw that breaks the camel's back.

Complexity theories argue against the linear, deterministic, patterned, universalizable, stable, atomized, modernistic, objective, mechanist, controlled, closed systems of law-like behaviour which may be operating in the laboratory but which do not operate in the social world of education. These features of complexity theories seriously undermine the value of experiments and positivist research in education (e.g. Waldrop, 1992; Lewin, 1993).

Complexity theory replaces these with an emphasis on networks, linkages, holism, feedback, relationships and interactivity in context (Byrne and Callaghan, 2014), emergence, dynamical systems, self-organization and an open system (rather than the closed world of the experimental laboratory). Even if one could conduct an experiment, its applicability to ongoing, emerging, interactive, relational, open situations, in practice, is limited (Morrison, 2001). It is misconceived to hold variables constant in a dynamical, evolving, fluid, open situation. What is measured is history.

Complexity theory challenges randomized controlled trials – the 'gold standard' of research. Classical experimental methods, abiding by the need for replicability and predictability, may not be particularly fruitful since, in complex phenomena, results are never clearly replicable or predictable: As Heraclitus noted, we never jump into the same river twice. Complexity theory suggests that educational research should concern itself with: (a) how multivalency and non-linearity feature in education; (b) how voluntarism and determinism, intentionality, agency and structure, life-world and system, divergence and convergence interact in learning (Morrison, 2002a, 2005); (c) how to both use, but transcend, simple causality in understanding the processes of education (Morrison, 2012); (d) how viewing a system holistically, as having its own ecology of multiple interacting elements, is more powerful than an atomized approach. Complexity theory suggests that phenomena must be looked at holistically; to atomize phenomena into measurable variables and then to focus only on certain of these is to miss synergy, the dynamic interaction of several parts (Morrison, 2008) and the significance of the whole. Measurement, however acute, may tell us little of value about a phenomenon; one can measure every observable variable of a person to an infinitesimal degree, but his/her nature, what makes him/her who he or she is, eludes atomization and measurement.

These should merge, so that in complexity theory the unit of analysis becomes a web, network or ecosystem (Capra, 1996, p. 301; Morrison, 2012), focused on, and arising from, a specific topic or centre of interest (a 'strange attractor'). Individuals, families, students, classes, schools, communities and societies exist in symbiosis; complexity theory tells us that their relationships are necessary, not contingent, and analytic, not synthetic. This is a challenging prospect for educational research, and complexity offers considerable leverage into understanding societal, community, individual and institutional change theory (Radford, 2006; Morrison, 2008); it provides the nexus between macro- and micro-research in understanding and promoting change.

In addressing holism, complexity theory suggests the need for case study methodology, narrative approaches, action research and participatory forms of research, premised in many ways on interactionist, qualitative accounts, i.e. looking at situations through the eyes of as many participants or stakeholders as

possible (e.g. Byrne and Callaghan, 2014; Boulton *et al*., 2015). This enables multiple causality, multiple perspectives and multiple effects to be charted (Morrison, 2012). Self-organization, a key feature of complexity theory, argues for participatory, collaborative and multi-perspectival approaches to educational research. This is not to deny 'outsider' research; it is to suggest that, if it is conducted, outsider research has to take in as many perspectives as possible.

In educational research terms, complexity theory stands against methodologies based on linear views of causality, arguing for multiple causality, multi-directional causes and effects and networks of causes (Morrison, 2012) at a host of different levels and in a range of diverse ways. No longer can one be certain that a simple cause brings a simple or single effect, or that a single effect is the result of a single cause, or that the location of causes will be in single fields only, or that the location of effects will be in a limited number of fields (Morrison, 2009, 2012). Researching causality becomes a search for networked, multi-causality and multi-stranded causality (Morrison, 2012).

Complexity theory not only questions the values of positivist research and experimentation, but it also underlines the importance of educational research to catch the deliberate, intentional, agentic actions of participants and to adopt interactionist and constructivist perspectives. (In this respect it has sympathies, perhaps, with posthumanism, though it is a very different animal from posthumanism.) Kuhn (2007, pp. 172–3) sets out a series of axioms for complexity-based research: (a) reality is dynamic, emergent and self-organizing, requiring multiple perspectives to be addressed (see also Medd, 2002); (b) the relationship between the knower and the known is, itself, dynamic, emergent and self-organizing; (c) hypotheses for research must relate to time and context (cf. Medd, 2002; Radford, 2006); (d) it is impossible to distinguish cause from effect, as entities are mutually shaping and influencing (co-evolution); (d) inquiry is not value-free.

Addressing complexity theory's argument for self-organization, the call is for the teacher-as-researcher movement to be celebrated, and complexity theory suggests that research in education could concern itself with the symbiosis of internal and external researchers and research partnerships. Just as complexity theory suggests that there are multiple views of reality, so this accords not only with the need to catch several perspectives on a situation (using multi-methods), but resonates with those tenets of critical research which argue for different voices, views and interpretations to be heard, incorporated and understood respectively. Heterogeneity is the watchword.

Complexity theory provides not only a powerful challenge to conventional approaches to educational research, but it suggests both a substantive agenda and also a set of methodologies, arguing for methodological, paradigmatic and theoretical pluralism. For example, Byrne and Callaghan (2014) and Boulton *et al*. (2015) suggest that research should study the processes of emergence over time and critical incidents in evolving situations. In addressing holism, complexity theory suggests the need for case study methodology, qualitative research and participatory, multi-perspectival and collaborative (self-organized), partnership-based forms of research, premised on interactionist, qualitative and interpretive accounts (e.g. Lewin and Regine, 2000).

1.18 Conclusion

This chapter has argued that planning and conducting educational research cannot follow simple recipes but is a complex, deliberative and iterative process in which ontological and epistemological matters have to be considered and in which many different kinds of understanding feature. In addressing this, the chapter has introduced several paradigms and their possible contribution to educational research, including: positivism, post-positivism, post-structuralism, postmodernism and complexity theory. It has commented on different views of social reality and a range of approaches to understanding that reality: deductive and inductive; empirical and rationalist; nomothetic and idiographic; subjective and objective; the scientific method; and alternatives in naturalistic, interpretive, phenomenological, interactionist and constructionist approaches.

The argument through the chapter has suggested that foundationalism and the quest for absolute knowledge in educational research is questionable. In this it has indicated the expanding range of approaches, of which, for example, postmodernism, post-structuralism and complexity theory are examples. Complexity theory challenges conceptions of simple cause-and-effect, experimental approaches to research and it advocates attention to context and holism in educational research.

In recognizing the many and expanding number of paradigms and approaches to educational research, the chapter has argued for methodological, paradigmatic and theoretical pluralism, indeed mixed methods (Chapter 2). These set the ground for the many approaches, designs, methodologies and methods set out in the remainder of the book. Simple recipe-following is out, and deliberation, fitness for purpose and fitness of purpose are key watchwords here.

The companion website to the book provides additional material and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: www.routledge.com/cw/cohen.

Notes

1 We are not here recommending, nor would we wish to encourage, exclusive dependence on rationally derived and scientifically provable knowledge for the conduct of education – even if this were possible. There is a rich fund of traditional and cultural wisdom in teaching (as in other spheres of life) which we would ignore to our detriment. What we are suggesting, however, is that total dependence on the latter has tended in the past to lead to an impasse, and that for further development and greater understanding to be achieved education must needs resort to the methods of science and research.

2 A classic statement opposing this particular view of science is that of Kuhn (1962), *The Structure of Scientific Revolutions*. Kuhn's book, acknowledged as an intellectual tour de force, makes the point that science is not the systematic accumulation of knowledge as presented in textbooks; that it is a far less rational exercise than generally imagined. In effect, 'it is a series of peaceful interludes punctuated by intellectually violent revolutions … in each of which one conceptual world view is replaced by another'.

3 The formulation of scientific method outlined earlier has come in for strong and sustained criticism. Mishler (1990), for example, describes it as a 'storybook image of science', out of tune with the actual practices of working scientists who turn out to resemble craftspersons rather than logicians. By craftspersons, Mishler is at pains to stress that competence depends upon 'apprenticeship training, continued practice and experienced-based, contextual knowledge of the specific methods applicable to a phenomenon of interest rather than an abstract "logic of discovery" and application of formal "rules" '. The knowledge base of scientific research, Mishler contends, is largely tacit and unexplicated; moreover, scientists learn it through a process of socialization into a 'particular form of life'. The discovery, testing and validation of findings is embedded in cultural and linguistic practices and experimental scientists proceed in pragmatic ways, learning from their errors and failures, adapting procedures to their local contexts, making decisions on the basis of their accumulated experiences. See, for example, Mishler (1990).

4 Investigating social episodes involves analysing the accounts of what is happening from the points of view of the actors and the participant spectator(s)/investigator(s). This is said to yield three main kinds of interlocking material: images of the self and others, definitions of situations, and rules for the proper development of the action. See Harré (1976).

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. In addition there is further information on complexity theory. These resources can be found online at **www.routledge.com/cw/cohen**.

Mixed methods research

CHAPTER 2

This chapter introduces:

- definitions of mixed methods research
- why use mixed methods research
- the foundations of mixed methods research
- paradigms and the commensurability problem in mixed methods research
- working with mixed methods approaches
- mixed methods designs and data
- reliability and validity in mixed methods research
- mixed methods research questions
- sampling in mixed methods research
- mixed methods data analysis
- timing and writing up the data analysis in mixed methods research
- stages in mixed methods research

2.1 Introduction

When we look at a phenomenon, do we suddenly don a quantitative hat, or a qualitative hat? Surely not. In viewing our world we naturally integrate rather than separate; we use all the means and data at our disposal to understand a situation. We use mixed methods to find out about something. So it can be in educational research. Mixed methods research (MMR) is not new (Denscombe, 2014, p. 159), but its new-found ascendancy and prominence, and indeed its title, have captured the world (cf. de Lisle, 2011). Claims made for MMR are not modest. The rise of MMR has been meteoric to the extent that it has been called the 'third methodological movement' (Johnson *et al*., 2007; Teddlie and Tashakkori, 2009), the 'third research paradigm' (Johnson and Onwuegbuzie, 2004; Johnson *et al*., 2007, p. 112; Denscombe, 2008) and the 'third path' (Gorard and Taylor, 2004), whilst Fetters and Freshwater (2015) suggest that the synergy of quantitative plus qualitative offers more than the individual components ('1 + 1 = 3' (p. 116)).

The 'paradigm wars' (Gage, 1989), in which one stood by one's allegiances to quantitative or qualitative methodologies, and which sanctioned the rise of qualitative methods and the partial eclipse of solely numerical methods (Denzin, 2008, p. 316), have given way to MMR (Gorard and Taylor, 2004; Gorard and Smith, 2006; Teddlie and Tashakkori, 2009). This recognizes that there is a need for greater rapprochement and less confrontational approaches to be adopted between different research paradigms (Denzin, 2008, p. 322), greater convergence between the two (Brannen, 2005), and a greater dialogue to be engaged between them and their proponents.

The placement of this chapter on MMR after the opening chapter in this book is deliberate, to acknowledge that, for many writers, MMR has its own paradigm, its own foundational views on social reality and research, its own ontology and epistemology, its own axiologies and methodologies. MMR already has a major place in research. It constitutes an approach, a methodology and a view of designs and methods (which we also set out in this chapter for the sake of fidelity to the principle of pragmatism that underlines MMR as well as for the sake of coherence and practical implications). The argument that we raise in this chapter is that, by virtue of its theoretical roots in pragmatism, its ontology and epistemology, its axiological premises, it is well located in Part 1. We also recognize that the later parts of this chapter could also sit comfortably in Parts 2 and 3, but this would be to fragment unnecessarily the discussion of MMR and lose the coherence to which MMR stakes an important claim.

The attention given to MMR is evidenced in the *Journal of Mixed Methods Research*, the *International Journal of Multiple Research Approaches*, an exponential increase in the number of key texts in the field and the launching of the Mixed Methods International Research Association (http://mmira.wildapricot.org).

MMR recognizes, and works with, the fact that the world is not exclusively quantitative or quantitative; it is not an either/or world, but a mixed world, even though the researcher may find that the research has a predominant disposition to, or requirement for, numbers or qualitative data. We see the world in multiple ways, some of which may or may not agree with each other. MMR encourages us not only to look at the world in different ways but to share those multiple, different

views in making sense of the world, discussing our views and values in it.

MMR not only relates to data collection, but concerns philosophical bases of research, paradigms which guide research and assumptions which inform the design and conduct of research. Creswell and Plano Clark (2011) observe that MMR brings together quantitative and qualitative data in a single research study or series of research studies (p. 5), the intention of which is to give a greater understanding of the topic or problem in question than either a quantitative or qualitative approach on its own would provide.

MMR focuses on collecting, analysing and mixing both quantitative and qualitative data in a single study or series of studies. Its central premise is that the use of quantitative and qualitative approaches, in combination, provides a better understanding of research problems and questions than either approach on its own. This is, in part, because research problems are not exclusively quantitative or qualitative, hence using only one kind of data (quantitative or qualitative), one methodology, one paradigm, one way of looking at the problem or one way of conducting the research, may not do justice to the issue in question (cf. Creswell and Plano Clark, 2011, p. 10; Creswell, 2012, p. 535). Further, a piece of research may have more than one phase, and MMR may take place both within and across phases. However, MMR is not only about data types; its reach extends much further, into ways of viewing the world, ontologies, epistemologies, axiologies, methodologies and a range of other areas which are introduced in this chapter.

2.2 What is mixed methods research?

Mixed methods research defies simple or single definitions, as the following references indicate.

Creswell and Plano Clark (2011, p. 4) offer an introductory definition in suggesting that MMR typifies research undertaken by one or more researchers which combines various elements of both quantitative and qualitative approaches (e.g. with regard to perspectives, data collection and data analysis) to research, together with the nature of the inferences made from the research (p. 4), the purposes of which are to give a richer and more reliable understanding (broader and deeper) of a phenomenon than a single approach would yield. Leech and Onwuegbuzie (2009, p. 265) suggest that conducting MMR involves data collection (both quantitative and qualitative), analysis and interpretation of studies that, singly or together, address a particular phenomenon.

However, MMR is not confined simply to methods, nor to methodology; rather it has a much wider embrace. MMR has many different definitions (Tashakkori and Teddlie, 2003). Johnson *et al*. (2007, pp. 119–21) give nineteen definitions that vary according to what is being mixed, where and when the mixing takes place, the breadth and scope of the mixing, the reasons for the mixing, and the orientation of the research. Greene (2008, p. 20) suggests that a mixed method way of thinking recognizes that there are many legitimate approaches to social research and that, as a corollary, a single approach on its own will only yield a partial understanding of the phenomenon being investigated.

As an example of its definitional pluralism, Tashakkori and Teddlie (2003) indicate that varieties of meanings of MMR lie in six major domains: (a) basic definitions; (b) utility of MMR; (c) paradigmatic foundations of MMR; (d) design issues; (e) drawing inferences; and (f) logistical issues in conducting MMR. Teddlie and Tashakkori (2006, 2009) set out seven dimensions in organizing different views of MMR:

- the number of methodological approaches used;
- the number of strands or phases in the research;
- the type of implementation process in the research;
- the stage(s) at which the integration of approaches occur(s);
- the priority given to one or more methodological approaches (e.g. quantitative over qualitative or vice versa, or of equal emphasis;
- the purpose and function of the research study;
- the theoretical perspective(s) in the research.

Creswell and Tashakkori (2007) set out four different realms of MMR which address what is being mixed: (a) methods (quantitative and qualitative methods for the research and data types); (b) methodologies (mixed methods as a distinct methodology that integrates world views, research questions, methods, inferences and conclusions); (c) paradigms (philosophical foundations and world views of, and underpinning, MMR); and (d) practice (mixed methods procedures in research). Clearly MMR operates at all stages and levels of research.

Greene (2008, pp. 8–10) organized discussion of MMR into four domains:

- philosophical assumptions and stances (assumptions about ontology – the nature of the world – and epistemology – how we understand and research the world, and the warrants we use in validating our understanding);

- inquiry logics (e.g. purposes and research questions, designs, methodologies of research, sampling, data collection and analysis, reporting and writing);
- guidelines for practice (how to mix methods in empirical research and in the study of phenomena);
- socio-political commitment (what and whose interests, purposes and political stances are being served).

Hesse-Biber and Johnson (2013) note that MMR applies to different paradigms, axiologies, stakeholders, levels of analysis (micro, meso, macro) and research cultures and practices (p. 103), recognizing that it is the research question that is central and critical in the design of the MMR and that research problems often require plural methodologies, cross-disciplinary approaches and multiple philosophical perspectives.

A mixed methods approach can apply to all the stages and areas of research: philosophical foundations and paradigms; ontologies, epistemologies, axiologies; methodology, research questions and design; instrumentation, sampling, validity, reliability, data collection; data analysis and interpretation; reporting; and outcomes and uses of the research (cf. Creswell and Tashakkori, 2007; Bergman, 2011a). This echoes Yin (2006, p. 42), who argues that the stronger the mix of methods and their integration at all stages, the stronger the benefit of mixed methods approaches (p. 46).

Clearly, even at the definitional and scoping stages, challenges are raised concerning what MMR is, how it can be conceptualized and organized, what it comprises and how it is conducted.

2.3 Why use mixed methods research?

It is claimed that MMR enables a more comprehensive and complete understanding of phenomena to be obtained than single methods approaches and answers complex research questions more meaningfully, combining particularity with generality, 'patterned regularity' with 'contextual complexity', insider *and* outsider perspectives (*emic* and *etic* research), focusing on the whole *and* its constituent parts, and the causes of effects (discussed in Chapter 6). Creswell and Plano Clark (2011) note that MMR can yield insights into, and explanations of, the processes at work in a phenomenon and the multiple views of the phenomenon (p. 61), thereby increasing the usefulness and credibility of the results found, indeed affording the opportunity for unexpected results to be found.

Denscombe (2014, p. 147) suggests that MMR can provide a more complete picture of the phenomenon under study than would be yielded by a single approach, thereby overcoming the weaknesses and biases of single approaches (the benefits of 'complementarity' and 'completeness' (Creswell and Plano Clark, 2011, p. 61)). Denscombe (2014) also suggests that MMR can increase the accuracy of data and reliability through triangulation, reduce bias in the research, provide a 'practical, problem-driven approach to research' (p. 160) and enable compensation between strengths and weaknesses of research strategies.

Day and Sammons (2008) indicate how a mixed method approach can provide more nuanced and authentic accounts (than single methods approaches) of the complexities of phenomena under investigation. Greene (2005, p. 207) argues for a mixed methods approach that welcomes multiple methodological traditions, as these catch diversity and difference and are 'anchored in values of tolerance, acceptance, respect' and democracy (p. 208). Mertens (2007) and Greene (2008) argue that, in seeking social justice, MMR operates in a 'transformative paradigm' (see Chapter 3).

Care has to be taken to separate 'complementarity' from 'supplementarity' in MMR. Whilst 'complementarity' suggests that one method may make up for the shortcomings of another, 'supplementary' is simply additive (cf. Bergman, 2011a), and, in itself, is not a sufficient justification for MMR, as any addition would meet this requirement. The researcher has to decide whether one method is being used to complement or supplement the research. If it is the former, then what is absent that the complementarity must rectify, and if it is the latter, what is being added or supplemented that renders it important for such addition or supplementation to be included? Further, unless the research question or problem unequivocally requires MMR, it is for the researcher to demonstrate that MMR in principle is preferable to a mono-method approach (p. 274).

In considering whether or not to employ MMR, and in addressing fitness for purpose, researchers can ask:

- What is gained/lost by looking/not looking at the world in mixed ways, i.e. using/not using MMR in terms of philosophical foundations, paradigms, ontologies, epistemologies, axiologies, methodologies, designs, research questions, sampling, data types, instrumentation, data analysis, data interpretation, drawing conclusions and reporting?
- What does researching objectively and subjectively, scientifically and interpretively, quantitatively and qualitatively, by numbers and by qualitative approaches, tell us?

- What is it about a piece of research that requires MMR, such that not to use MMR is to diminish the quality, validity, reliability and utility of the research?

2.4 The foundations of mixed methods research

Paradigms and pragmatism

Mixed methods research has several foundations (cf. de Lisle, 2011, pp. 91–2). For example, quantitative approaches may have their roots in positivism, post-positivism and the scientific paradigm. Qualitative methods may have their roots in the interpretive paradigm. Transformative approaches may appeal to critical theory with its political and ideological agenda of empowerment, emancipation, equality and social justice. The foundations of MMR have multiple allegiances, and these allegiances determine and embrace world views (what the world is like and how to look at the world), ontologies (views of reality), epistemologies (ways of understanding, knowing about and researching that reality) and axiologies (values and value systems, e.g. value-free or value-laden research). These are brought together in different ways in different paradigms.

A paradigm, following Kuhn (1962), defines 'the set of practices that define a scientific discipline at any particular period of time' (p. 175): what is to be observed and scrutinized; the kinds of research questions to be asked and problems to be investigated; how to structure such research questions; what predictions can be made by the primary theory in that discipline; the ways of working; and how to interpret results. A paradigm embodies the values and beliefs of a group (in Kuhn's case it was scientists), such that one set of views and beliefs may be incommensurable with another, abiding by different philosophical assumptions, ontologies, epistemologies and axiologies. Mertens (2012) suggests that paradigms are 'philosophical frameworks that delineate assumptions about ethics, reality, knowledge, and systematic inquiry' (p. 256). Paradigms include how we look at the world, the conceptual frameworks in which we work in understanding the world, the community of scholars who are working within that framework and who define what counts as worthwhile knowledge and appropriate methodology in it, how we research the world, what the key concepts are, what counts as relevant knowledge and how we validate and consider that knowledge.

Given that a 'paradigm' embraces a 'world view', to define a paradigm in terms of quantitative, qualitative or mixed methods is misleading, as these refer largely or only to kinds of data (Biesta, 2010a), and a paradigm has a much wider embrace than this which includes a world view, an epistemological stance, shared beliefs and model examples (Freshwater and Cahill, 2013, p. 50). MMR concerns not only mixing data but mixing paradigms, ontologies, epistemologies and axiologies in order to give a fair, rounded picture of the phenomenon under investigation.

Creswell and Plano Clark (2011, p. 40) identify four paradigms or world views (see also Chapter 1):

- Post-positivism (quantitative research), in which emphasis is placed on the identification of causality and its effects, focusing on variables and their manipulation (e.g. isolation and control of variables in a reductionist world), careful observation and measurement, and hypothesis testing in a world characterized by a singular view of reality and in which the researcher imposes the research on the phenomenon (i.e. top-down).
- Constructivism (qualitative research), in which the objective of the research is to understand a phenomenon as it is seen and interpreted by the participants themselves, individually (e.g. Piagetian constructivism) or socially (e.g. Vygotskyian constructivism) in a world characterized by a multiple view of reality and in which the researcher works with the world as it is construed by its participants (i.e. bottom-up).
- Participatory/transformative (qualitative research), in which the research has a deliberate agenda of seeking to improve the situation of its participants, focusing, thereby, on issues of: agentic control of one's life; power, empowerment, social justice, marginalization and oppression; voice and action, all in a world characterized by a political, negotiated view of reality and in which the researcher works collaboratively with participants to improve the life situation of disempowered groups and individuals.
- Pragmatism (quantitative and qualitative), in which the research focuses on framing and answering the research question or problem, which is eclectic in its designs, methods of data collection and analysis, driven by fitness for purpose and employing quantitative and qualitative data as relevant, i.e. as long as they 'work' – succeed – in answering the research question or problem, and in which the researcher employs both inductive and deductive reasoning to investigate the multiple, plural views of the problem and the research question.

Mertens (2012) identifies three paradigms in MMR: 'dialectical pluralism', lodged between constructivism and post-positivism (p. 256); 'pragmatism' and the

'transformative' paradigm (p. 256). She argues that these paradigms in MMR have 'different sets of philosophical assumptions' (p. 256), though it is questionable where the incommensurability question is actually answered here, as incommensurability does not evaporate by making different data types available in a single piece of research. This rehearses the differences between mixing data, methods and world views in MMR.

Morgan (2007) argues against the use of the term 'paradigm' in MMR, suggesting its replacement by 'approach', particularly in his advocacy of the pragmatic approach. In MMR, methodological pluralism is the order of the day as this enables errors in single approaches to be identified and rectified (Johnson *et al*., 2007, p. 116). It also enables meanings in data to be probed, corroborated and triangulated, rich(er) data to be gathered and new modes of thinking to emerge where paradoxes between two individual data sources are found (p. 115; Sechrest and Sidana, 1995). For example, one can adopt a constructivist approach in developing a research problem or question, and then adopt a pragmatic, post-positivist or transformative paradigm for investigating it (Flick *et al.*, 2012). At issue here is whether commencing in one paradigm frames a research question or problem in a way that would be different if one had commenced in a different paradigm. A paradigm affects how we think about a problem or issue (Mertens and Hesse-Biber, 2012).

Much MMR works beyond quantitative and qualitative exclusivity or affiliation, and instead operates in a 'pragmatist paradigm' (Onwuegbuzie and Leech, 2005; Ercikan and Roth, 2006; Johnson *et al*., 2007, p. 113; Teddlie and Tashakkori, 2009, p. 4; Gorard, 2012, p. 8) which draws on, and integrates, both numeric and narrative approaches and data, quantitative and qualitative methods where relevant, to meet the needs of the research rather than the allegiances or preferences of the researcher, and in order to answer research questions fully. Whereas post-positivist approaches are premised on scientific, objectivist ontologies (how we construe reality) and epistemologies (how we understand, come to know about or research reality), and whereas interpretive approaches are premised on humanistic and existential ontologies and epistemologies, by contrast, MMR is premised on pragmatist ontologies and epistemologies.

Quantitative approaches are not all of one kind, and neither are all qualitative approaches. In this respect, Onwuegbuzie and Leech (2005, p. 377) argue that not all quantitative approaches are positivist and not all qualitative approaches are hermeneutic. For example, quantitative approaches can catch opinions, perceptions, probabilistic causality and process approaches (e.g. structured observation), and qualitative approaches can feature in experiments, identifying causality, surveys and patterns of, and trends in, data (e.g. Miles and Huberman, 1984, 1994).

Onwuegbuzie and Leech (2005, p. 376) argue that MMR recognizes similarities between different philosophies and epistemologies (in quantitative and qualitative traditions), rather than the differences that keep them apart, and that there are far more similarities than differences between the two approaches, as both use observational data, both describe data, and construct explanations and speculations about the reasons why observed outcomes are as they are (p. 379). Both concern corroboration and elaboration; both complement each other and identify important conflicts, where they arise, between findings from the two kinds of data (cf. Brannen, 2005, p. 176).

Hammersley (2013) suggests that the terms 'quantitative research' and 'qualitative research' are no longer useful categories (p. 99), as there are major variants of each, and he suggests, rather, that in conducting research it is preferable to use a range of strategies that lend themselves to 'research practice' (p. 99). Methodological puritanism should give way to methodological pragmatism in addressing research questions (cf. Caracelli and Greene, 1993; Greene, 2008; Creswell, 2009).

A commonly given basis of MMR is pragmatism. This is loosely interpreted to be 'what works', i.e. if the methods of research and the data collected – be they numerical or qualitative – address the research purposes, problems or questions then they are acceptable. In other words, the research is driven by the research question. Biesta (2012) contrasts a pragmatic approach with a principled approach (p. 147), though this is contestable, as pragmatism is no less a principle or a philosophical position than, say, post-positivism or constructivism. The principle underpinning pragmatism is that thought should lead to action, to prediction and problem solving.

Pragmatists such as James, Peirce and Dewey consider thought to be an instrument or tool for accurate prediction, problem solving and action, i.e. philosophy is not merely a contemplative exercise but is judged by its practical outcomes, success in practice, ability to solve problems and the everyday use-value of philosophizing. What is 'true' and what is valuable is 'what works'. As Ulysse and Lukenchuk (2013, p. 18) remark, in pragmatism one is less concerned with the truth or falsehood of an idea and more concerned with whether the idea can make a difference (they quote William James's comment that pragmatism concerns its 'cash value'). Similarly they note that Peirce's

pragmatism concerned less a theory of truth and more whether a solution can be found to a problem.

Pragmatism is essentially practical rather than idealistic; it is 'practice-driven' (Denscombe, 2008, p. 280). It argues that there may be both singular and multiple versions of the truth and reality, sometimes subjective and sometimes objective, sometimes scientific and sometimes humanistic. It is a matter-of-fact approach to life, oriented to the solution of practical problems in the practical world. It prefers utility, practical consequences, outcomes and heurism over the pursuit of a single, particular kind of accuracy in representing 'reality'. Rather than engaging in the debate over qualitative or quantitative affiliations, it gets straight down to the business of judging research by whether it has found out what the researcher want to know, regardless of whether the data and methodologies are quantitative or qualitative (Feilzer, 2010, p. 14).

In pragmatism, what something 'means' is manifested in its practical, observable consequences and success in practices, with its links to experience, rather than, for example, abstract theory with little practical import, or ideology, or dogmatic adherence to a particular value system or epistemology. Theories are to be judged by their practical utility rather than being ends in themselves; they are instruments for coping with, understanding and living with 'reality'. Hence a 'good' theory pulls its weight in its practical utility; values and beliefs denote rules for action.

Working in this vein argues against any privileged, distinctive method of enquiry; 'what works' is what helps us to understand, research and solve a problem. Our frames of reference, conceptual schemes, categories for understanding the world, are not immutable or eternal, but are our creations, our artefacts, useful insofar as they solve practical problems. Which frameworks, categories, theories, conceptual schemes and ways of viewing a problem we use are decided by their practical utility and applicability in solving a particular problem. Knowledge and action are closely connected and mutually informing.

Clearly pragmatism is no less value-based than other 'principles'; it is simply that its values differ from others. Pragmatism adopts a methodologically eclectic, pluralist approach to research, drawing on positivist, post-positivist and interpretive epistemologies based on the criteria of fitness for purpose and applicability, and regarding 'reality' as both objective and socially constructed (Johnson and Onwuegbuzie, 2004). No longer is one a slave to methodological loyalty and a particular academic community or social context (Oakley, 1999). Denscombe (2008) argues for the mixed methods paradigm to be defined in terms of a new 'community of practice' of those like-minded researchers who adopt the principles of MMR; regarding MMR in terms of a 'community of practice' respects the pragmatic underpinning of this approach.

Pragmatism suggests that 'what works' to answer the research questions is the most useful approach to the investigation, be it a combination of experiments, case studies, surveys or whatever, as such combinations enhance the quality of the research (e.g. Suter, 2005). Indeed Chatterji (2004) argues that mixed methods are unavoidable if one wishes to discover 'what works'. Pragmatism is not an 'anything goes', sloppy, unprincipled approach; it has its own standards of rigour, and these are that the research must answer the research questions and 'deliver' useful, practicable, reliable and valid answers to questions put by the research.

Paradigms and the commensurability problem in mixed methods research

Mixed methods research has to grapple with the issue of 'commensurability': is it possible to mix methods which have distinct and incompatible roots and views of the world, and how we should research and understand it, what should we look for and look at, and how should we make sense of the world?

Whether paradigms are or are not incommensurable, whether they can coexist alongside each other or can be integrated, is an immense open, philosophical question. Bergman (2011a, 2011b) comments that the recourse to pragmatism is no solution to, or resolution of, the incompatibility problem; it still exists and will continue to exist as it is illogical to try to seek coherence of such incoherence in a single research design (2011a, p. 101) (see also Denzin, 2012), even if it 'works' in practice. Hammersley (2013) argues that quantitative and qualitative approaches are irreconcilable as their rationales are very different (p. 97), such that mixing quantitative and qualitative methods means, in effect, 'abandoning key assumptions' of qualitative research (p. 97). Indeed Borge (2012, p. 15) notes that there are times when, rather than trying to mix methods, it may be helpful to have different specialisms and division of labour in quantitative and qualitative terms: we need specialists to give us expert advice on particular aspects of a phenomenon.

Biesta (2012, p. 148) identifies seven levels of 'mixing', and he raises challenging questions for those working with MMR:

1 '*Ontologies*', questioning whether and how it is possible combine different ontologies (e.g. views of the nature of reality).

2 '*Epistemologies*', questioning whether and how it is possible to combine different epistemologies (ways of knowing).
3 '*Research purposes*', questioning whether and how it is possible to combine the wish to have research which seeks causal explanations with that which seeks understanding and interpretation.
4 '*Practical orientation*', questioning whether and how it is possible for research to be directed both towards producing 'solutions, techniques and technologies' (p. 148) and towards developing 'critical understanding'.
5 '*Designs*', questioning whether and how it is possible to combine interventionist designs, such as experiments, with non-interventionist designs, such as naturalistic research.
6 '*Data*', questioning whether and how it is possible to combine text and numbers.
7 '*Methods*', questioning whether and how it is possible to combine different methods of collecting and/or analysing data.

Biesta's view goes to the heart of the dilemma of MMR, questioning whether a piece of research can genuinely 'mix' different elements (as in mixing water and milk to form a new liquid) or simply combine them but keep them separate (as in combining the separate pieces of a jigsaw to make a complete picture). We return to 'commensurability' and incommensurability later in this chapter.

Bergman (2011a) notes that even the term 'mixing' is inappropriate because one cannot mix that which cannot be mixed, and he argues that MMR designs are unable to bridge incompatible ontological, epistemological and axiological positions (p. 273). How, he asks, can one combine a subjectivist foundation with an objectivist one, or research that separates the researcher from the research with that which binds them together? He argues that more suitable terms than 'mixed' might be 'blended', 'meshed' and 'combined' (p. 272). Similarly Creswell and Plano Clark (2011, p. 277) comment that mixed methods differ from multi-methods, in that multi-methods do not necessarily imply that they will be mixed. In terms of educational research this suggests the need to identify the benefits of each approach (e.g. quantitative and qualitative) in terms of the overall research purpose, problem or question.

Consider the analogy: was it possible for scientists to work in two distinct paradigms – the geocentric view which put the Earth at the centre of the universe (a Ptolemaic model) or a heliocentric view with the sun at the centre (the Copernican view)? Surely these two are fundamentally incompatible? Applying this analogy to MMR calls into question whether, in fact, it is fitting to call MMR a paradigm at all. For example, in what sense can I combine an atheistic view of the world with a theistic view of the world and then call this a new paradigm? The two have fundamentally different and irreconcilable starting positions, rationales, values, foundations and ways of looking at the world, and to bring them together under a convenient label of a 'paradigm' is a misnomer; it does not 'mix' them at all, it just puts them side by side and draws on each as appropriate in answering a research question or problem. In this instance we have two paradigms, not one. Maybe MMR is just a convenient shorthand for something that we understand but which has different and incompatible premises, and which is not actually a single paradigm, or, more generously, is a paradigm based on compatibility – each party living in comfort alongside the other – rather than mixing, i.e. a marriage rather than a metamorphosis into a single organism.

Putting together quantitative and qualitative designs and data may be difficult, as the two may be incommensurate in terms of the paradigms, ontologies, epistemologies, methodologies, axiologies, data types, etc. The analogy may be made with trying to mix oil and water, which stay separate, rather than milk and water, which mix. Recognizing such differences may not be a problem as, together, complementarily, they can yield a complete picture of the phenomenon in question. Oil and water may not mix but they give more than oil alone or water alone.

Further, neither is quantitative nor qualitative research all of one type. For example, not all quantitative research is large scale and not all qualitative research is small scale (cf. Miles and Huberman, 1984, 1994). 'Quantitative' and 'qualitative' are umbrella terms, each covering a multitude of research types. Hence, in designing MMR, specificity is necessary about what kind of research is planned with respect to the quantitative and qualitative elements.

In relation to the issues of the incommensurability of paradigms (Howe, 1988; Denzin, 2008; Creswell, 2009, p. 102; Trifonas, 2009, p. 297), MMR argues for their compatibility, or at least their ability to live alongside each other and to work together to solve a research problem. These same authors suggest the power of integrating different approaches, ways of viewing a problem, and types of data in conducting research, induction and deduction in answering research questions, in strengthening the inferences that can be made from research and data and in generating theory. Indeed Reams and Twale (2008, p. 133) argue that mixed methods are necessary and important in addressing information and perspectives, and that they 'increase

corroboration of the data, and render less biased and more accurate conclusions'. Maybe that leaves behind the problem of whether MMR constitutes a paradigm, whether quantitative and qualitative approaches can be brought into a single overarching paradigm, or whether each is incommensurable with the other. In other words, whether or not we recognize commensurability and incommensurability actually doesn't matter that much, if at all, in the 'real world' of practical utility in MMR.

Researchers need not become mired in the paradigm debate; as long as we know what we are dealing with in MMR then this may suffice. Mertens and Hesse-Biber (2012) suggest it is time to move beyond the commensurability/incommensurability question (p. 75). We still have not resolved the incompatibility thesis, but that does not mean that we are unable to move forward in MMR (Bergman, 2011b) or to conduct MMR research.

2.5 Working with mixed methods approaches

There are no blueprints for how to work with MMR; each piece of research is unique and the researcher has to decide how to design and implement the research, based on its own purposes, foci, merits and characteristics. What follows, then, are considerations in coming to these decisions in terms of design issues, research questions, sampling, data collection and analysis, and writing up the data analysis. We leave behind the issue of paradigms and their commensurability, and move to planning 'what works', as this accords with the pragmatic roots of MMR.

Mixed methods research designs and data

A research design is the plan for, and foundations of, approaching, operationalizing and investigating the research problem or issue; setting out the approach, theory/ies and methodology/ies to be employed; the types of data required, how they will be collected (instrumentation) and from whom (the population and/or sample); how the data will be analysed, interpreted and reported; the warrants to be adduced to defend the conclusions drawn and the degree of trust that can be placed in the validity and reliability of each element of the research; and the sequence of the research.

In MMR the kinds and methods of research and its several stages or phases are driven by the research questions or research problem, with 'fitness for purpose' as a guiding principle. There must be a clear matching of the research question to the research problem and to the methods used for answering that research inquiry. For MMR this means providing a reasoned and reasonable justification for mixing whatever elements of the research design are, indeed, to be mixed (e.g. world views, views of reality, paradigms, rationales, theories, methodologies and approaches, data types and instrumentation, sampling, data analysis, interpretation and reporting, types of validity and reliability), stages and phases of the research, conclusions, outcomes and consequences of the research.

In approaching MMR designs, key decisions have to be taken on several issues (cf. Teddlie and Tashakkori, 2009, p. 141; Creswell and Plano Clark, 2011, pp. 64–7):

- Why used a mixed methods approach? What will a mixed methods approach provide that a non-mixed methods approach does not?
- What, actually, will be mixed, and why, for example, paradigms, ontologies, epistemologies, theories and theoretical frameworks, designs, research purposes and questions, methodologies, populations and samples, data types, data-collection instruments and their contents, data analysis, interpretation and reporting?
- Why, where, at what level(s), in what areas and how will this 'mixing' occur, how will it be done, adhering to what principles, procedures and processes?
- When, where, why and how will the designs and data be mixed, merged, integrated, connected, adhering to what principles, procedures and processes, and how will the quantitative designs and data relate to qualitative designs and data, and vice versa? How and why will one design be embedded in another?
- What methodologies will be used, where, when, why and how?
- How many strands, levels, stages and phases will there be in the research, and where, how and why do quantitative and qualitative approaches feature in these? What will be the relative priority accorded to the quantitative and qualitative strands, for example, will they have equal priority/importance, will one take priority over the other, and, if so, at which stages or phases of the research, and why?
- What will be the level and type of interaction between the quantitative and qualitative strands of the research, for example, will they be independent, separate, integrated, combined, parallel, interactive?
- What will be the timing and/or sequence of the quantitative and qualitative strands in the research, for example, will they be concurrent/parallel and/or sequential in a time sequence within and between phases, and why?
- What ethical issues does MMR present?

Teddlie and Tashakkori (2009) suggest different designs in MMR. '*Parallel mixed designs*' (p. 26) (also termed 'concurrent designs') are those in which both qualitative and quantitative approaches run simultaneously but independently in addressing research questions, akin to the familiar notion of triangulation of method, theory, methodologies, investigators, perspectives and data, discussed later in this book. '*Sequential mixed designs*' (p. 26) are those in which one or other of quantitative and qualitative approaches run one after the other, as the research requires, and in which one strand of the research or research approach determines the subsequent strand or approach and in which the major findings from all strands are subsequently synthesized. '*Quasi-mixed designs*' (p. 142) are those in which both quantitative and qualitative data are gathered but which are not integrated in answering a particular research question, i.e. quantitative data might answer one research question and qualitative data another research question, even though both research questions are included in the same piece of research. '*Conversion mixed designs*' (p. 151) are those in which data are transformed (qualitative to quantitative and vice versa, e.g. in a parallel mixed design) (the issues of quantitizing qualitative research and qualitizing quantitative research are discussed below). '*Multilevel mixed designs*' (in parallel or sequential research designs) (p. 151) (also termed 'hierarchical' research designs) are those where different types of data (both quantitative and qualitative) are integrated and/or used at different levels of the research (e.g. student, class, school, district, region), for instance numerical data may be used at one level (students) and qualitative data used at another level (school). '*Fully integrated mixed designs*' (p. 151) are those in which mixed methods are used at each and all stages (perhaps iteratively: where one stage influences the next) and levels of the research.

Creswell and Plano Clark (2011) identify six MMR designs in which timing and sequence feature strongly. They contend that there must be a valid warrant or justification for the sequence and design chosen, and note that samples and sample sizes may vary with each kind of data and at different stages of the research. Their *convergent parallel design* (pp. 69–79) has both quantitative and qualitative data which are collected independently and in parallel with each other, and then they converge, yielding triangulation of data and offering complementary data on the question, problem, issue or topic in question. Quantitative and qualitative data are collected and analysed separately and then put together, for example they may be compared and contrasted, looking for similarity, difference and complementarity. The overall, combined or integrated results are reported.

In an *explanatory sequential design* (pp. 82–4), quantitative data are usually collected first, followed by qualitative data to explain the quantitative data. It is important for the researcher to identify which parts of the quantitative data need to be explained and how they can be explained (and with which sample(s)).

Their *exploratory sequential design* (pp. 86–7) reverses the sequence of data collection in the explanatory sequential design; qualitative data are usually collected first (typically with a small sample), with quantitative data from a larger sample used to generalize the findings.

Their *embedded design* (pp. 90–2) recognizes that each research question requires both quantitative and qualitative data, and qualitative data may be added to, embedded in or supplemented by quantitative data (e.g. in an experiment) or vice versa (e.g. a case study) in this design. In the former (the experiment), the qualitative data may be used to explain and interpret the quantitative data, whilst in the latter (the case study) the quantitative data may provide additional, more generalized data on the case (e.g. frequencies). The authors note that one type of data tends to have priority over another in this design: for example, qualitative data may be embedded within a largely quantitative study or quantitative data may feature within a mainly qualitative study. The authors also note that quantitative and qualitative data tend to be kept separate. It is important to decide when, and in what sequence, to collect the data: for example, concurrently and/or sequentially. In discussing an embedded design, Creswell and Plano Clark introduce a widely used notation:

QUAN = Quantitative data which have priority over qualitative data
Quan = Quantitative data which are subordinate to qualitative data
QUAL = Qualitative data which have priority over quantitative data
Qual = Qualitative data which are subordinate to quantitative data

They also introduce other symbols in outlining notation in designs (pp. 108–10):

+ (the methods – quantitative and qualitative – occur simultaneously);
() (one method is embedded within another);
→ (a linear sequence, where one stage informs the next or is kept separate);
→← (the methods are used recursively);

[] (mixed methods operate within a single study or a series of studies);
= (the outcome of the mixing).

For example, a case study may be characterized as '(QUAL and Quan)', whereas an experiment may be characterized as '(QUAN and Qual)'. The authors indicate the sequence of the quantitative/qualitative methodology, data collection and analysis by a simple arrow (→). We outline some conceptual MMR designs using these (Figure 2.1).

In their *transformative design* (pp. 96–7), as in critical theory, there is an explicitly political or ideological, social intention or agenda, to advance the social justice for the group or groups under study. In this collaborative, participatory type of research, the authors suggest

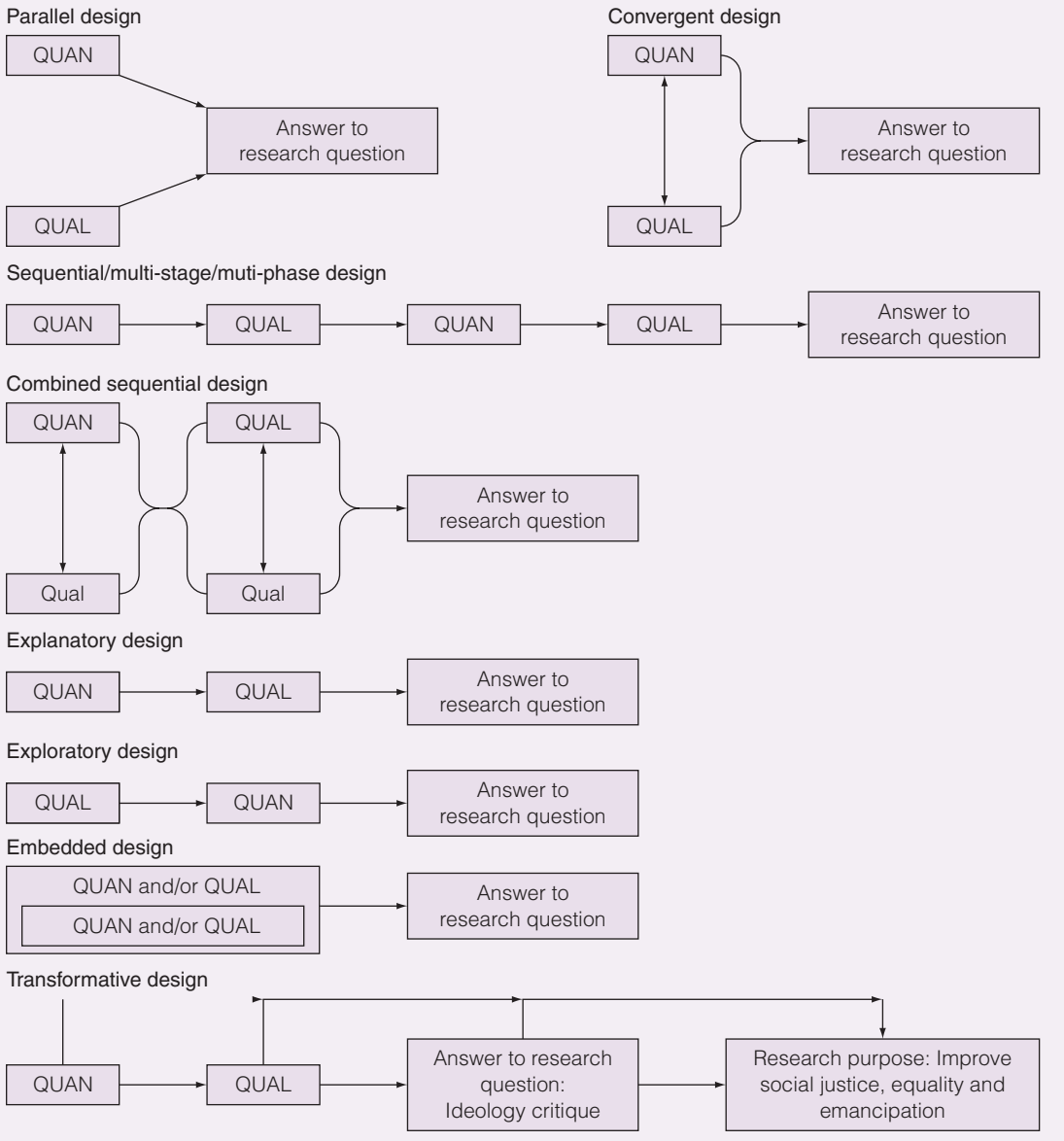


FIGURE 2.1 Mixed methods research typologies

that quantitative data precede qualitative data. However, in this design it is less the data types and sequence that are important as the overall purpose of the research, i.e. the research has a political/ideological agenda (whether this is the legitimate concern of researchers is another matter, for example Hammersley (2014, chapter 3) questions whether researchers should concern themselves with what uses are made of their data and, rather, should concentrate on ensuring that their research is conducted rigorously and without bias). As we argue in Chapter 3, the methodologies of research in the critical theory approaches are ideology critique and action research (Carr and Kemmis, 1986).

Finally, in their *multi-phase design* (Creswell and Plano Clark, 2011, pp. 100–11) the quantitative and qualitative data can be concurrent and/or sequential, depending on the phase of the research in which they are being used. At issue here is the need to identify the key phases of the research as it unfolds, and then decide which kind of data are needed in each phase. The point here is that the progress of the research is incremental and cumulative: one phase builds on, and is informed and influenced by, the preceding phase in addressing the overall purposes of the research. Hence the decision of which kinds of data are required at each stage is an iterative one, and it is important that each phase of the research is connected clearly. The authors comment that this kind of research is often characterized as a series of 'mini-studies' leading towards the overall answer to the research question or problem.

These are suggested models; clearly there are very many variants on these designs, as there may be enormous variety of: timing; number of stages/phases; sequence; data types in the sequence and within each stage; the priority/weights given to data types; interaction/independence of data (de Lisle (2011) provides a useful summary of these). It is for each research study to plan its own design. Even though mixed methods may be used, in some research the numerical approach may predominate – with its own sampling implications – whilst in others qualitative data may predominate, with an emphasis on purposive and non-probability sampling (cf. Teddlie and Yu, 2007, p. 85).

The designs set out above are not exhaustive, nor are they discrete, nor do they indicate the levels (other than data) at which the quantitative and qualitative aspects operate (e.g. paradigms; world views; ontologies; epistemologies; axiologies; methodologies; instrumentation; sampling; data types, collection, analysis, interpretation, reporting etc.). There is no single methodological approach in MMR (Hesse-Biber and Johnson, 2013). Rather, the typologies set out above are ideal types and typifications for the sake of heuristic clarity, designed to alert researchers to different kinds of MMR. It is for each research study to plan its own design. The design types set out above identify key issues to be addressed (e.g. Ivankova *et al*., 2006, pp. 9–11; Greene, 2008, pp. 14–17):

- *The paradigm dimension*: which paradigms are operating in the research, and why? For example, Creswell and Plano Clark (2011) align post-positivism with quantitative research, constructivism with qualitative research, transformative research with the transformative design, and pragmatism with those designs which are directed to answering the research question or problem regardless of which data types are used. This is not to argue that research is, or must be, paradigm-driven; rather it is to say that different kinds of design may be present within an overall study, and that the logic of each design type should be integrated into the overall logic of the entire study.
- *The methodology dimension*: which methodologies/approaches will be used (e.g. survey; experiment; case study; ethnography, interpretive and interactionist approaches; action research; historical study), which will impact on the research design, sampling, instrumentation, data analysis, ethics?
- *The time dimension*: when and where will the quantitative and qualitative elements be present in the study – in what sequence and/or concurrence or simultaneity? Should the quantitative and qualitative data be analysed together or separately?
- *The priority dimension*: which and what has priority (if any), where and when – quantitative and qualitative (e.g. paradigms, methodologies, data types, data analysis)?
- *The relationship dimension*: will the research types and data types be independent, interactive, complementary, additional to each other? What are the relationships between different types of data at different points in the research, both within-phase/within-stage and cross-phase/cross-stage?
- *Integration*: where and when – at which stages – and why do the integration of quantitative and qualitative methods and data occur?
- *Independence, the obverse of integration*: where, when and why will methods and data be kept concurrent, separate, interactive or independent?
- *Differentiation*: will mixed methods and data be used to address the same issue or different issues?
- *Matching*: which kinds of data are required for which stages of the research?
- *Issues in question*: around what issues do the mixed methods occur, for example, at the levels of

constructs, variables, research questions, purposes of the research?
- *Transformative intention*: does the research have an explicitly political agenda?
- *Scope*: does the mixing of methods occur within a single study or across more than one study in a set of coordinated studies within a single programme of research?
- *Strands*: how many different strands are mixed in the study (Greene, 2008, p. 14)?

In reality, the cleanness of the designs set out above may not catch the reality of conducting research, which, in many cases, is characterized by multiple iterations, modifications and emergence rather than a pre-figured design. Indeed Creswell and Plano Clark (2011, p. 105) note that designs may be fixed from the very beginning or may emerge as the study unfolds. For example, there is no golden rule which states that such-and-such a design or data type should precede or succeed another or that data can only be analysed or mixed at such-and-such a point or points in time; the decision is taken on fitness for research purpose and fitness for research question. We present different designs in Figure 2.1.

Kettley (2012) questions the usefulness of delineating an unending host of different designs of MMR at all, deeming such attempts to be 'unproductive labour' (p. 85). This is uncharitable, as such delineation can stimulate and clarify, without shackling, the deliberative process needed in deciding what is to be the appropriate design for a given piece of research. Typologies have heuristic value, and, indeed may indicate the relative importance of the quantitative or qualitative elements (Denscombe, 2014, p. 151). Pluralism and fitness for purpose, rather than slavish adherence to a single pre-fixed design, are the order of the day. Indeed research designs may change and emerge over the course of a study; the process is an emergent part of a dynamical system. Each design is different and must be decided by the research in hand.

There must be a defensible reason for mixing data types. For example, qualitative data may be used to develop instruments (e.g. a pre-pilot); to understand the context of research and the participants in it; to validate the quantitative data; to understand participants' views of the research and what is being studied; to gain feedback on an intervention; to identify the effects and impact of an intervention and its unanticipated effects and risks; to understand the processes of an intervention and the changes in participants over time; to identify intervening factors; to explain cause and effect; to explain, understand and triangulate the quantitative data.

On the other hand, quantitative data may be used for generalizing the outcomes of research or an intervention; providing 'hard' data; measuring effects of an intervention; refining data-collection instruments (e.g. removing unreliable items or items which too strongly correlate with other items); gaining an overall picture and patterns of response; identifying, measuring and modelling correlations and relationships, differences, key underlying factors; and suggesting cause and effect.

The mixed methods researcher has the same battery of instruments available for data collection as for mono-methods research. These are set out in the several chapters of this book. Of concern here are the implications of the 'mixed' nature of MMR for mixing data. Whilst this is taken up in the prior discussion of MMR designs, at issue here is whether, how and where to mix data, the warrants that attach to each, and ensuring the validity and reliability of the resultant mix. Underpinning this is the point that a genuine 'mix' means fidelity not only to the different nature and warrants of quantitative and qualitative data but also to the fact that both types must be demonstrably relevant to answering a given research question and must be fit for purpose.

Timing is an important dimension of the research design in respect of data types in MMR. Qualitative data may be useful *before* an experiment/trial commences, for example for: ensuring that the research meets a need; instrument development; gaining informed consent; understanding more about the participants; and gaining baseline data. This differs from the use of qualitative data *during* an experiment/trial, which here may be for: data validation and triangulation; impact analysis; gaining participants' perceptions of and opinions on what is occurring; understanding what is happening and why; identifying resource needs; identifying emerging issues and factors affecting the process. In turn, this differs from the use of qualitative data *after* an experiment/trial, which may be to gain participants' perceptions of and opinions and feedback on what had happened; to determine outcomes, effects and impact; to suggest explanations of or reasons for what had happened; and to compare before-and-after situations.

MMR addresses both the 'what' (numerical and quantitative data) and 'how or why' (qualitative) types of research questions. This is particularly important if the intention of the researcher is really to understand different explanations of outcomes. For example, let us say that the researcher has found that a hundred people decide that schools are like prisons. This might be an interesting finding in itself, but it might be that forty of the respondents thought they were like prisons because

they restricted students' freedom and had very harsh, controlling discipline. Twenty respondents might say that schools were like prisons because they were overcrowded; fifteen might say that schools were like prisons because the food was awful; ten might say that schools were like prisons because there was a lot of violence and bullying; ten might say schools were like prisons because they were 'total institutions' (Goffman, 1968); and another five might say that schools were like prisons because students had an easy life as long as they obeyed the rules. Here the reasons given for the simple statistic are very different from each other, and it is here that qualitative data can shed a lot of useful light on a simple statistic.

Reliability and validity in mixed methods research

Including quantitative and qualitative data may offer greater *reliability*. Within quantitative and qualitative approaches this includes a range of elements (see Chapter 14): for example, respondent validation, credibility of results, replicability, equivalence, stability, internal consistency and Cronbach alphas, dependability, credibility, accuracy, fidelity to context etc. These ensure reliability *within* each approach (quantitative and qualitative). Further, reliability-as-triangulation includes *between methods* approaches: for example, instruments, data types, researchers, time, participants, perspectives (people and approaches: objective and subjective, inductive and deductive (Morgan, 2007; Torrance, 2012); theories; methodologies; paradigms; axiologies; designs). Denscombe (2014, pp. 154–5) suggests that triangulation can be: (a) *methodological* (between methods), enabling researchers to study a phenomenon from a variety of perspectives and using dissimilar methods; (b) *methodological* (within methods), i.e. those methods which are similar to each other; (c) *data triangulation* (using contrasting sources of information, e.g. from different people, at different times, in different locations); (d) *investigator* (different researchers); and (e) *theory* (different theoretical positions).

Combining quantitative and qualitative data may also strengthen the *validity* of the research and the inferences that can be drawn from it in: the rigour of the design and its fitness for purpose in meeting the research purposes and research questions; methodological rigour; consistency of findings and conclusions with the evidence presented; defensible and credible inferences drawn; and the quality of the synthesis of data.

Validity *within* an approach is required, and Chapter 14 addresses this. Validity in quantitative and qualitative approaches have their own canons of rigour. In ensuring validity *between* approaches, Teddlie and Tashakkori (2009) argue that 'meta-inferences' assess the extent and degree to which the sets of inferences from quantitative and qualitative approaches are credible (cf. Ivankova, 2013), and that credible research requires such meta-inferences to be addressed and to be legitimate. Validity in MMR requires: designs that are appropriate for the research questions, methodologies and sampling; consistency with all the components of the study; procedures employed for analysing data to be appropriate to answer the research questions; and the different strands or elements of the MMR to be connected appropriately (Ivankova, 2013).

Ivankova (2013, p. 48) sets out a three-step process of validation of meta-inferences in MMR which employ a QUAN → QUAL design:

Step 1: Using a systematic process for selecting which participants to include in a qualitative follow-up;
Step 2: Elaborating, following up on and probing unexpected results from the quantitative data and their analysis;
Step 3: Observing and reporting on interactions between quantitative and qualitative strands of the study.

At issue here is the point that reliability and validity within each element/stage/data type of the research must be complemented by reliability and validity when combining the different elements/stages/data types of the research. We refer the reader here to Chapter 14, which includes more discussion of reliability and validity in mixed methods research.

Mixed methods research questions

In MMR the research is driven by the research questions (which require both quantitative and qualitative data to answer them). Greene (2008, p. 13) comments that methodology follows from the purposes and questions in the research rather than vice versa, and that different kinds of MMR designs follow from different kinds of research purposes: for example, hypothesis testing, understanding, explanation, democratization (see the discussion of critical theory in Chapter 3). Such purposes can adopt probability and non-probability samples (see Chapter 12), multiple instruments for data collection, and a range of data analysis methods, both numerical and qualitative.

In considering whether to adopt an MMR study, it is important for researchers to look at the research question or problem and ask themselves whether a single method on its own is appropriate or sufficient to answer

or address this respectively. If the answer is 'yes', then why consider MMR? If the answer is 'no', then what is needed from the quantitative and qualitative elements in order to answer the question or problem, and where should they be mixed or kept separate?

Tashakkori and Creswell (2007, p. 207) write that 'a strong mixed methods study starts with a strong mixed methods research question', and they suggest that such a question could ask 'what and how' or 'what and why' (p. 207), i.e. the research question, rather than requiring *only* numerical or qualitative data, is a 'hybrid' (p. 208). The research question, in fact, might be broken down into separate sub-questions, each of which could be either quantitative or qualitative, as in 'parallel' or concurrent mixed methods designs (see above) or in 'sequential mixed designs' (see above), but which converge into a combined, integrated answer to the research question (see also Chapter 10). Bryman (2007a, p. 13) goes further, to suggest not only that qualitative and quantitative data must be mutually informing, but that the research design itself has to be set up in a way that ensures that integration will take place, i.e. so that it is not biased to, say, a numerical survey.

Such research questions could be, for example: 'What are the problems of staff turnover in inner city schools, and why do they occur?' Here qualitative data might provide an indication of the problems and a range of reasons for these, whilst numerical data might provide an indication of the extent of the problems. Here qualitative data subsequently might be 'quantitized' into the numbers of responses expressing given reasons, or the quantitative data subsequently might be 'qualitized' in a narrative case study.

Sampling in mixed methods research

The material here does not rehearse the chapter on sampling, and readers are referred to Chapter 12. Here we confine ourselves to issues of sampling in MMR. Teddlie and Yu (2007) and Teddlie and Tashakkori (2009, pp. 180–1) indicate that it is commonplace for MMR to use more than one kind of sample (probability, non-probability) and to use samples of different sizes, scope and types (cases: people; materials: written, oral observational; other elements in social situations: locations, times, events etc.) within the same piece of research.

In MMR, sampling in quantitative approaches should address issues and criteria that are relevant to such quantitative approaches: for example, sampling strategy, probability and non-probability sampling, sample size calculation (with references to confidence intervals, confidence levels, sampling error and statistical power), choice of sample, representativeness, and access to the sample. In other words, sampling in quantitative approaches should abide by the canons of sampling principles for quantitative studies. This is not to say naively that samples in quantitative approaches should be large; they may be large, small and/or variable, depending on fitness for purpose, research questions and research design.

Similarly, qualitative approaches should abide by the canons of sampling in qualitative research, which address similar issues as quantitative approaches but have different decisions made on, or answers given to, those issues, for example on sampling strategy, purposive sampling, representativeness, access, size. This is not to say naively that samples in qualitative research should be small; they may be small, large and/or variable.

However, given the specifically *mixed* nature of MMR, consideration should be given to the implications of this for sampling, for example:

- What sampling strategies will be used for which elements of the research and will the same or different samples be used in both the quantitative and qualitative elements, for example, to ensure 'carry-through' and consistency of people, as having different samples may bring inconsistencies and undue divergence (Ivankova, 2013, p. 42)?
- Will the qualitative sample be drawn from the sample used in the quantitative element (i.e. some 'carry-through' of the sample, with the qualitative sample becoming, in effect, a sample of the quantitative sample), and will the qualitative sample include, but add to, the sample used in the quantitative element? If the qualitative sample is drawn from the quantitative sample, i.e. a sample of the sample, how will the qualitative sample be chosen?
- Will the quantitative sample be drawn from the sample used in the qualitative element (i.e. some 'carry-through' of the sample, with the quantitative sample becoming, in effect, a sample of the qualitative sample), and will the quantitative sample include, but add to, the sample used in the qualitative element? If the quantitative sample is drawn from the qualitative sample, i.e. a sample of the sample, how will the quantitative sample be chosen?
- At what point in the research will the samples be drawn, i.e. when will you decide whom the sample will comprise?
- Will the samples for the quantitative and qualitative elements be of the same or different sizes?
- Will the same or different samples be used for the same research question(s) and issues under study?

In some MMR studies it may be possible to decide the exact members of the sample(s) in advance of commencing the entire research, whereas in others it may be that choosing members of the sample may not be possible until a particular stage of the research has taken place. However, this does not mean that the *principles* for the sampling at different stages or for different elements of the research may not be decided in advance, only that the actual members for every stage of element may be unknown in advance.

For example, it may be that an initial quantitative survey in an MMR study may yield 'average' responses together with outliers, and that the qualitative element of the same overall study is designed to conduct follow-up interviews with some respondents whose responses were 'average' and others whose responses were outliers, i.e. to include in the qualitative sample members whose responses to the quantitative survey showed maximum variation. We do not know in advance who they will be, but we know the *principle* on which the qualitative sample will be selected.

An example of this is given by Ivankova (2013). She reports an MMR study of an online research methods training course which commenced with a quantitative survey (N = 119), and, following the statistical analysis of the numerical data, a sample of those from the quantitative survey was drawn for follow-up qualitative telephone interviews (N = 13). The sample for the qualitative interviews was purposive, chosen to be able to help the explanation and elaboration of the quantitative data (including unexpected results), and was based on the principles of seeking to reduce potential bias and socially desirable responses.

As another example, an MMR study might commence with a small-scale qualitative, exploratory set of interviews which raise issues to be included in a larger-scale quantitative survey which will require a random stratified sample, stratified according to characteristics that emerge in the initial interviews. Again, we do not know in advance who will be targeted for inclusion in the quantitative survey, but we know the *principle* on which the quantitative sample will be selected.

A major decision will concern whether to have entirely independent samples in the quantitative and qualitative approaches – different members in each sample – or whether to have any overlap of members. Decisions on this matter may depend on fitness for purpose. For example, Monteiro and Morrison (2014) report a study of undergraduate collaborative blended learning in which an initial large-scale survey was conducted on a population of students in one university, followed by a targeted quasi-experiment with a sample of students from one year-group of this population, gathering both quantitative and qualitative data from a purposive sample drawn from high-, medium- and low-performing students, using classroom observations, learning logs and interviews. This 'carry-through' of students for the quantitative and qualitative elements of the research enabled comparisons to be made between the survey data and the qualitative data, using the large-scale survey as a context in which the quasi-experiment was embedded.

Teddlie and Tashakkori (2009, pp. 185–91) provide a useful overview of different mixed methods sampling (see also Chapter 12). This includes *parallel mixed methods sampling*, *sequential mixed methods sampling*, *multilevel mixed methods sampling*, *stratified purposive sampling*, *purposeful random sampling* and *nested sampling designs*. Each of these, with examples, is addressed in Chapter 12, and we refer readers to that chapter. In the same chapter we note that the sampling strategy should derive logically from the research questions or hypotheses being investigated/tested. It should also be faithful to the assumptions on which the sampling strategies are based (e.g. random allocation, even distributions of characteristics in the population etc.). Each sample should generate sufficient qualitative and quantitative data in order to answer the research questions and enable clear inferences to be drawn from both the numerical and qualitative data. Sampling, of course, must abide by ethical principles and be practicable and efficient. Researchers should also consider whether the data will enable generalizability of the results to be addressed and to whom the results are generalizable. Further, the sampling should be reported at a level of detail that will enable other researchers to understand it and perhaps use it in the future.

Mixed methods data analysis

It is a truism to say that analysing quantitative and qualitative data must be faithful to the canons of quantitative and qualitative analysis respectively, and these are addressed in different chapters of this book (Part 5). These operate when treating quantitative and qualitative data separately. However, MMR asks for the integration of, and connection between, quantitative and qualitative data.

Quantitative and qualitative data can be analysed separately and independently, as, for example, in parallel or sequential designs (e.g. quantitative to qualitative or vice versa), and they can also be mutually informing. For example, Ivankova (2013) reports how, after she had conducted her quantitative data analysis and then proceeded to her qualitative data analysis, her qualitative data analysis suggested that she needed to go back and conduct further statistical analysis of her numerical

data. The process of data analysis in MMR is iterative, not necessarily a once-and-for-all event for each element or stage of the research. The researcher will need to decide:

- the purposes of data analysis both during and after the research process;
- which tools to use for analysis (e.g. numbers, words, graphics), what kind of analysis is most suitable for what kinds of data, what to look for in different kinds of data (e.g. do the different kinds of data focus on the same issue or different issues?), how to present different kinds of data analysis (e.g. in prose, tables, graphics), how to analyse the quantitative and qualitative data (see Part 5), and how to apply 'constant comparison' (see Chapter 37) to compare them, looking for similarities, differences, contrasts, additions, refinements, extensions, contradictions, mutual reinforcements, supplements, complements etc.;
- whether and why to analyse quantitative and qualitative data separately, independently or together, i.e. what, if any, is the relationship between the data types and their analysis?;
- the sequence and timing of the data analysis: when to analyse each kind of data, whether, why – and, if appropriate, how – to use the analysis of one kind of data to inform subsequent data collection and analysis and whether, when, where, why and how to relate, connect, merge and/or integrate data and data types;
- whether, where, how and why to quantitize qualitative data and to qualitize quantitative data, how to combine, compare and represent different types of data in answering a research question (e.g. analyse quantitative data and then qualitative data, or vice versa, and then draw key messages/themes from them together);
- which data in the data analysis have greater priority, and why, and how to represent and address this;
- what to do if the results from the analysis of one kind of data contradict, support, refine, qualify, extend those of another kind of data, what to do if re-analysis of earlier data is required, and what to do if inadequate, insufficient or weak data are found;
- how to combine data if they derive from different sampling strategies and different, unequal sample sizes, types and people.

Some kinds of research require 'progressive focusing' (Parlett and Hamilton, 1976), in which a study commences with a broad field of view and analyses data on this broad picture in order to identify key features. These features are then investigated further, in closer detail, moving from a wide view to a much narrower, focused set of issues. In MMR, for example, this lends itself to the analysis of large-scale quantitative data identifying patterns and key features, similarities and differences, which are then explored, for example in focus groups, observational data or semi-structured qualitative interviews. The point here is that one set of data analysis both precedes and informs what comes next.

MMR can combine data types (numerical and qualitative) in answering research questions and also convert data (Bazeley, 2006, p. 66). Caracelli and Greene (1993) suggest four strategies for integrating and converting data in MMR (see also Creswell and Plano Clark, 2011, p. 213): (a) data transformation (discussed below); (b) typology development (where classifications from one set or type of data are applied to the other set or type of data); (c) extreme case analysis (where outliers found in one set of data are explored using different data and methods); and (d) data consolidation/merging (where new variables are created by merging data).

'Data conversion' ('transformation') (Teddlie and Tashakkori, 2009, p. 27) is where qualitative data are 'quantitized' (converted into numbers, typically nominal or ordinal; see Chapter 38) (e.g. Miles and Huberman, 1994). This can be done, for example, by giving frequency counts of certain responses, codes, data or themes in order to establish regularities or peculiarities, or rating scales of intensity of those responses, data, codes or themes (Sandelowski *et al*., 2009, p. 210; Teddlie and Tashakkori, 2009, p. 269). Software can also assist the researcher in providing frequency counts of qualitative data (e.g. Bazeley, 2006). 'Data conversion' can also take place where numerical data are 'qualitized' (converted into narratives and then analysed using qualitative data analysis processes).

It is misguided to imagine that different types of data can somehow be truly mixed, as if their different nature simply disappears. MMR recognizes that data are different, but that is not the issue. Rather, the issue is how they can be combined, related and merged. In this, the answer is both simple and difficult: be guided by the research question. It is the logic of the research question that impacts on the data analysis. In answering the research question, both quantitative and qualitative data might be adduced, each calls on its own warrants and claims to validity and reliability. The differences are intrinsic; oil is not water, and that is the beauty of each of them, but that does not mean we cannot draw on both in addressing an issue.

Timing and writing up the data analysis in mixed methods research

Bryman (2007a, p. 8) indicates a signal feature of MMR that distinguishes it from the simple usage of quantitative and qualitative research separately within a single piece of research; here mixed methods researchers write up their research in 'such a way that the quantitative and qualitative components are mutually illuminating'. This criterion of 'mutually illuminating' not only argues for the fully integrated mixed design but it also calls for research purposes and questions to *require* such integration, i.e. that the research question cannot be answered sufficiently by drawing only on one or the other of quantitative or qualitative methods, but that it requires both types of data.

The researcher is faced with several decisions in writing up the data analysis: for example, when to conduct and/or write up the data analysis (e.g. during or after the research, at the end of each stage or phase of the research in a sequential study); how to organize the presentation/write-up to answer each research question (e.g. by sample and sub-sample, individuals, theme, topic, research question, instrument, data type, stage/phase of the research etc.; see Part 5); whether one data type or stage of the research influences another data type or stage of the research (e.g. do the findings from quantitative data influence the qualitative data at that stage, or are they kept independent; whether the findings from one stage (e.g. quantitative stage) influence what happens in the next, qualitative stage); and how to organize the write-up of the data analysis in each stage or phase.

A major question here is whether one stage of the research influences the subsequent stage, even if, within each stage, mixed methods are being used. For example, in an explanatory design the quantitative data might suggest areas that the subsequent qualitative data should explain; in an exploratory design the qualitative data might suggest areas to be explored in the subsequent quantitative data. In these instances the timing of the data analysis is critical, as it is impossible to proceed to the next stage until the preceding data analysis is completed.

In a *parallel design*, with quantitative and qualitative data kept separate until the point of convergence, it would seem appropriate to organize the writing-up of the data analysis by the research question. But then the researcher has to decide, when writing up the data analysis in answering the research question, whether to present the data analysis separately by data type (e.g. qualitative and quantitative), or by different themes in answering the research question (with relevant quantitative and qualitative data integrated in addressing each theme), or by sample/sub-sample or instrument.

In a *sequential design* (e.g. quantitative followed by qualitative) it might be more appropriate to organize the data analysis and write-up first by stage/phase of the research and then draw this all together at the end of the data analysis to answer the research question. At each phase the researcher faces a similar set of decisions as in a parallel design, i.e. how to organize the write-up of the data analysis: by sample and sub-sample, individuals, theme, topic, research question, instrument or data type.

In an *explanatory sequential design* the qualitative data collection may come after the quantitative data. Here, for clarity, it may be useful to follow the same sequence in presenting the data analysis, with the quantitative data preceding the qualitative data, followed by a section which draws together the two data types in answering the research question. In an exploratory sequential design the sequence is reversed, with the quantitative data collection coming after the qualitative data. Here, for clarity, it may be useful to follow the same sequence in presenting the data analysis, with the qualitative data preceding the quantitative data, followed by a section which draws together the two data types in answering the research question.

In an *embedded design* one kind of data is subordinate to, or embedded within, another major data type. In this situation the main data may be presented first, with the supplementary data ensuing. It may be that the write-up of the data analysis takes the form of a case study, in which the quantitative and qualitative data are integrated in a narrative that 'tells the story' of the case. This latter can also apply to transformative designs.

The above designs are only typologies. As mentioned earlier, there are no blueprints for how and when to conduct and write up the data analysis. Each piece of research suggests its own most suitable designs, and these may be iterative and emergent, with several stages which move from quantitative to qualitative data and vice versa and their consequent own suitable ways of presenting the data analysis and the timing of these. Fitness for purpose is complemented by the need for clarity, relevance and ease in understanding the data and how they answer the research question. Indeed, in many cases the text of the write-up is exactly that – a text – in which both numbers and words appear as appropriate.

Consider, for example, a case study of an intervention to improve school attendance. Here overall school figures on attendance and absence may be addressed at

the start of, or even before, the intervention. Quantitative and qualitative data may give rise to the research (e.g. frequency of absence from school), leading to qualitative and quantitative data from analysis of records, followed by analysis of further quantitative data, followed by exploratory interviews, followed by re-analysis of qualitative and qualitative data, and so on. Each stage of the research is driven by the data analysis at the preceding stage, and the researcher in this MMR design has to decide when is the appropriate time to conduct and use the data analysis. The logic of each stage of the design and the research question decides where, when and how to combine the quantitative and qualitative data, and indeed the overall write-up of the research may be a narrative which draws freely on both numbers and words.

2.6 Stages in mixed methods research

Creswell (2012, pp. 554–7) sets out a seven-step process in MMR planning and conduct:

Step 1: Determine whether a mixed methods study is practicable and feasible.
Step 2: Set out the rationale for mixing methods (justify the use of MMR and justify the model of MMR being used).
Step 3: Set out the data-collection strategy (consider the priority, sequence and kinds of qualitative and quantitative data required).
Step 4: Develop quantitative, qualitative and mixed methods questions.
Step 5: Collect quantitative and qualitative data.
Step 6: Analyse data separately, concurrently or both.
Step 7: Write the report as a one- or two-phase or a multi-phase study.

However, this overlooks a more exact indication of what is to be mixed. Hence we suggest a twelve-step process:

Step 1: Decide the purpose of the research.
Step 2: Decide the nature of the phenomenon or problem that you wish to research, such that MMR is the most appropriate approach.
Step 3: Decide the research questions, ensuring that they can only be answered fully by the provision and analysis of mixed data.
Step 4: Decide what is to be 'mixed' in the MMR: ontologies (views of reality); paradigms (world views, lenses through which to define the problem and how to consider the research, and commensurate ways of working in the research); epistemologies; axiologies; theories and theoretical frameworks; research designs; methodologies and approaches; data types; data-collection instruments and methods; sampling; data; data analysis, interpretation and reporting; types of validity, validation and reliability.
Step 5: Decide the stages and phases of the research, where the 'mixing' will occur in these stages/phases and which kinds of methodologies and data are pre-eminent at each stage or phase.
Step 6: Decide the data collection (quantitative and qualitative and their interrelations), what (kind of) data are required from whom, when and at what stage(s) and phase(s).
Step 7: Design the data-collection instruments and the sampling.
Step 8: Collect the data.
Step 9: Plan the data analysis including: the function of the data analysis (e.g. formative, summative, an ongoing record), which data have priority, when and where, the timing (e.g. ongoing, at the end of each phase, at the end of the entire project) and sequence of data analysis.
Step 10: Conduct the data analysis, being clear on which data, from whom, and when the data and their analysis will be mixed, related, kept separate, interactive, when the analysis will commence overall and by stage or phase.
Step 11: Decide how to organize and write the research report, for example, by phase, by data types, where to integrate data types, where to comment on the points in Step (4).
Step 12: Write the research report.

Clearly in a multi-phase research design several of these steps will be repeated, or the sequence altered (e.g. Step 9 may precede Step 8).

As can be seen here, the research question (Step 3), though it may drive the MMR, is itself the consequence of prior considerations (Steps 1 and 2), and MMR must be able to justify itself in terms of addressing these prior considerations. As Biesta (2012, p. 149) remarks, the research question, far from being the first step in the research, is itself the operationalized consequence of the research purposes and problems.

2.7 Conclusion

This chapter has suggested that MMR constitutes an important way of looking at the social and educational

world that is informed by a pragmatic paradigm of practicality in answering research purposes and research question – 'what works' in planning, conducting and reporting the research – which rests on a range of ontological, epistemological and axiological foundations. For many years pragmatism has emerged as a prevailing principle to guide researchers. In order to give coherence to the discussion, the chapter then moved from the material on paradigms, principles, ontologies and epistemologies, to a practical account of its implications for the practice of research, thereby embodying the 'practicality' spirit of pragmatism that underpins MMR. In this spirit the chapter discussed matters of research designs, research questions, sampling, methodologies, reliability and validity, data types, data collection and analysis, and reporting.

The chapter also raised some challenges for MMR, for example, whether it really constitutes a new paradigm and how it addresses the problem of commensurability and incommensurability of the paradigmatic roots that underpin quantitative and qualitative research. Further, on the one hand, the advocates of MMR hail it as an important approach that is rooted in pragmatism, which: (a) yields real answers to real questions; (b) is useful in the real world; (c) avoids mistaken allegiance to either quantitative or qualitative approaches on their own; (d) enables rich data to be gathered which address the triangulation that has been advocated in research for many years; (e) respects the mixed, messy real world; and (f) increases validity and reliability; in short, that 'delivers' 'what works'. MMR possesses the flexibility in usage that reflects the changing and integrated nature of the world and the phenomenon under study. Further, it draws on a variety of ways of working and methodologies of enquiry, ontology, epistemology and values. It is a way of thinking, in which researchers see the world as integrated and in which they have to approach research from a standpoint of integrated purposes and research questions. As has been argued in this chapter, MMR enters into all stages of the research process: (a) philosophical foundations, paradigms, ontologies, world views, epistemologies and axiologies; (b) research purposes and research questions; (c) research design, methodology, sampling, data types, instrumentation and data collection, validity and reliability; (d) data analysis; (e) data interpretation; (f) conclusions and reporting results.

On the other hand, MMR has been taking place for years, before it was given the cachet of a new paradigm; it is not unusual for different methods to be used at different stages of a piece of research or even at the same stage, or with different samples within a single piece of research. It does not really have the novelty that seems to be claimed for it. Further, underneath MMR are still existing quantitative and qualitative paradigms, and they are different in world views, ontologies, epistemologies and axiologies, so to mix them by bringing them under a single sobriquet of 'mixed methods research' may be a disingenuous sleight of hand. There is also the matter of the perceptions which reveal underlying sympathies to paradigms and/or views of combining research types: imagine that we mix water with wine; is the liquid which results from such mixing 'fortified water' or 'diluted wine' – strengthened or weakened?

Giddings (2006), Giddings and Grant (2007) and Hesse-Biber (2010) question whether there is suppressed, or covert, support for positivism or quantitative approaches residing within MMR. Further, can one call a paradigm new simply because it brings together two previous paradigms and makes a case for thinking in a mixed method way of answering research questions by different types of data? The jury is still out, though this book underlines the importance of combining methods where necessary and relevant in planning and doing research, and we return to MMR throughout the book, as an indication of its importance.

Denscombe (2014, p. 161) notes that MMR might entail increasing the time costs of the research and will require researchers who are skilled in more than one method. One can add to his point that there is an additional skill required in being able to combine methods. Further, MMR might give rise to problems if data from different methods do not corroborate each other, requiring the researcher to explore why this might be (de Lisle (2011, p. 106) notes that qualitative findings might provide contradictory rather than complementary data). MMR might misinterpret the philosophy of pragmatism to be expediency rather than principled action (e.g. 'anything goes') (Denscombe, 2014, p. 161).

In a wide-ranging review, Creswell (2011) identifies eleven key controversies in MMR:

1. What actually MMR is in a context of shifting and widening definitions of MMR (method, methodology, orientation, philosophy, world view, a way of seeing).
2. The usefulness of quantitative and qualitative descriptors (i.e. that the binary nature of these two terms does not hold in practice and is unnecessarily limiting).
3. Whether MMR is as new as some of its claimants might propose.

4 What really drives the interest in MMR (including the interests of funding agencies).
5 The relevance and usefulness of debates on paradigms and whether they can actually be mixed.
6 The putative privileging of post-positivism in MMR, and the consequent diminishing status of qualitative approaches, for example, in 'embedded' designs.
7 Whether there is a 'fixed discourse' in mixed methods, who controls it and whether mixed methods is becoming a new metanarrative.
8 Whether MMR should adopt a 'bilingual language' for its terms, i.e. whether a language should move beyond the vocabulary which might favour quantitative or qualitative approaches to a new, non-partisan glossary of terms.
9 The usefulness of a plethora of designs and typologies, which become confusing and betray the complexity of the phenomena under study.
10 Whether MMR is 'misappropriating' designs and methodologies from other fields of, and approaches to, research, and whether MMR might be 'a subordinate procedure within ethnography' (p. 280).
11 What the added value of MMR is, i.e. what it offers by way of understanding a research issue better than either quantitative or qualitative approaches alone offer.

These suggest that, though MMR has been around for decades, there are still many questions to be answered. Hesse-Biber and Johnson (2013) suggest that MMR still has 'gaps and opportunities', including, for example: ethical issues and team approaches in MMR; 'retooling' 'methods and traditions' whose origins lie in quantitative or qualitative research to bring them into MMR; implications of web-based developments; and big data and analytics for MMR.

Whilst there is a powerful case for MMR, the argument here has been that the researcher has to decide whether and how to use MMR, and that these decisions must be driven by fitness for purpose.

The companion website to the book provides PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: www.routledge.com/cw/cohen.

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Critical educational research

CHAPTER 3

This chapter sets out key features of critical theory as they apply to educational research, and then it links these to:

- critical theory and critical educational research
- participatory action research
- feminist theory
- value-neutrality in educational research

It recognizes that other approaches can be included under the umbrella of critical theory (e.g. post-colonial theory, queer theory), and, whilst the chapter includes a note on these, it does not develop them. Indeed critical theory embraces a range of other theories, for example, critical race theory, critical pedagogy, critical disability theory.

3.1 Critical theory and critical educational research

Positivist and interpretive paradigms are essentially concerned with understanding phenomena through two different lenses. Positivism strives for objectivity, measurability, predictability, controllability, patterning, the construction of laws and rules of behaviour, and the ascription of causality; interpretive paradigms strive to understand and interpret the world in terms of its actors. In the former, observed phenomena are important; in the latter, meanings and interpretations are paramount. Giddens (1976) describes this latter as a 'double hermeneutic', where people strive to interpret and operate in an already interpreted world; researchers have their own values, views and interpretations, and these affect their research, and, indeed, that which they are researching is a world in which other people act on their own interpretations and views.

It was suggested in Chapter 2 that mixed methods research has an affinity with equity, social justice and a 'transformative paradigm' (Mertens, 2007), and it is to this that we turn now. This paradigm of *critical educational research* regards the two previous paradigms of positivism and interpretivism as presenting incomplete accounts of social behaviour when they neglect the political and ideological contexts of educational research. Positivistic and interpretive paradigms are seen as preoccupied with technical and hermeneutic knowledge respectively (Grundy, 1987; Gage, 1989). The paradigm of critical educational research is influenced by the early work of Habermas and, to a lesser extent, his predecessors in the Frankfurt School, most notably Adorno, Marcuse, Horkheimer and Fromm. Here the expressed intention is deliberately political – the emancipation of individuals and groups in an egalitarian society.

Critical theory is explicitly prescriptive and normative, entailing a view of what behaviour in a social democracy *should* entail (Fay, 1987; Morrison, 1995a). Its intention is not merely to give an account of society and behaviour but to realize a society that is based on equality and democracy for all its members. Its purpose is not merely to understand situations and phenomena but to change them. In particular it seeks to emancipate the disempowered, to redress inequality and to promote individual freedoms within a democratic society. In doing so it focuses not only on individuals and groups, but also on society and its institutions and social arrangements, and it uses both evaluative and descriptive concepts (Hammersley, 2013, p. 30) such as exploitation, empowerment, class division, emancipation, justice, interests and suchlike, with the intention of bringing about specific political aims: equality, social justice, democracy, freedom from oppression and exploitation, and the transformation of society to an emancipated democracy within which people are empowered to take control over their own lives and life choices.

In this enterprise, critical theory identifies the 'false' or 'fragmented' consciousness (Eagleton, 1991) that has brought an individual or social group to relative powerlessness or, indeed, to power, and it questions the legitimacy of this. It holds up to the lights of legitimacy and equality issues of repression, voice, ideology, power, participation, representation, inclusion and interests. It argues that much behaviour (including research behaviour) is the outcome of particular illegitimate, dominatory and repressive factors, illegitimate in

the sense that they do not operate in the general interest – one person's or group's freedom and power is bought at the price of another's freedom and power. Hence critical theory seeks to uncover the *interests* at work in particular situations and to interrogate the legitimacy of those interests, identifying the extent to which they are legitimate in their service of equality and democracy. Its intention is *transformative*: to change society and individuals to social democracy. In this respect the purpose of critical educational research is intensely practical and political, to bring about a more just, egalitarian society in which individual and collective freedoms are practised, and to eradicate the exercise and effects of illegitimate power. The pedigree of critical theory in Marxism is not difficult to discern. For critical theorists, researchers can no longer claim neutrality and ideological or political innocence.

Critical theory and critical educational research have their substantive agenda: for example, examining and interrogating the relationships between school and society; how schools perpetuate or reduce inequality; the social construction of knowledge and curricula, who defines worthwhile knowledge; what ideological interests schools serve and how this reproduces inequality in society; how power is produced and reproduced through education; whose interests are served by education and how legitimate these are (e.g. rich, white, middle-class males rather than poor, non-white females); how different groups in society fare (e.g. by social class, gender, race, physical features, ethnicity, disability, sexuality) and how political goals might be achieved; in other words, the emancipation of all social groups regardless of social class, gender, race, physical features, ethnicity, disability, sexuality etc. Researchers, then, have an obligation to promote certain political views and to achieve certain political goals.

The significance of critical theory for research is immense, for it suggests that much social research is comparatively trivial in that it *accepts* rather than *questions* given agendas for research, compounded by the funding for research, which underlines the political dimension of research sponsorship (discussed later) (e.g. Norris, 1990). Critical theorists would argue that the positivist and interpretive paradigms are essentially technicist, seeking to understand and render more efficient an existing situation, rather than to question or transform it.

Critical approaches recognize that peoples, social groups, institutions and societies operate on the basis of 'interests' which are allied to ideologies and values. Habermas's early work (1972) offers a useful tripartite conceptualization of 'interests'. He suggests that knowledge – and hence research knowledge – serves different interests. Interests, he argues, are socially constructed, and are 'knowledge-constitutive', because they shape and determine what counts as the objects and types of knowledge. Interests have an ideological function (Morrison, 1995a), for example, a 'technical interest' (discussed below) can have the effect of keeping the empowered in their empowered position and the disempowered in their powerlessness, reinforcing and perpetuating the status quo. An 'emancipatory interest' (discussed below) threatens the status quo. In this view, knowledge – and research knowledge – is not neutral (see also Mannheim, 1936). What counts as worthwhile knowledge is determined by the social and positional power of the advocates of that knowledge. The link here between objects of study and communities of scholars echoes Kuhn's (1962) notions of paradigms and paradigm shifts, discussed in Chapters 1 and 2. Knowledge and definitions of knowledge reflect the interests of the community of scholars who operate in particular paradigms. Habermas (1972) constructs the definition of worthwhile knowledge and modes of understanding around three cognitive interests:

i prediction and control;
ii understanding and interpretation;
iii emancipation and freedom.

He names these the '*technical*', '*practical*' and '*emancipatory*' interests respectively. The technical interest characterizes the scientific, positivist method, with its emphasis on laws, rules, prediction and control of behaviour, with passive research objects: instrumental knowledge. The practical interest, an attenuation of the positivism of the scientific method, is exemplified in the hermeneutic, interpretive methodologies outlined in qualitative approaches. Here research methodologies seek to clarify, understand and interpret the communications of 'speaking and acting subjects' (Habermas, 1974, p. 8).

Hermeneutics focuses on interaction and language; it seeks to understand situations through the eyes of the participants, echoing the *verstehen* approaches of Weber (Ringer, 1997) and premised on the view that reality is socially constructed (Berger and Luckmann, 1967). Indeed Habermas (1988, p. 12) suggests that sociology must understand social facts in their cultural significance and as socially determined. Hermeneutics involves recapturing the *meanings of* interacting others, recovering and reconstructing the *intentions* of the other actors in a situation. Such an enterprise involves the analysis of *meaning in a social context* (Held, 1980). Gadamer (1975, p. 273) argues that the hermeneutic sciences (e.g. qualitative approaches) involve

the *fusion of horizons* between participants. Meanings rather than phenomena take on significance here.

The emancipatory interest subsumes the previous two paradigms; it requires them but goes beyond them (Habermas, 1972, p. 211). It is concerned with *praxis* – action that is informed by reflection with the aim of emancipation. The twin intentions of this interest are to expose the operation of power and to bring about social justice, as domination and repression act to prevent the full existential realization of individual and social freedoms (Habermas, 1979, p. 14). The task of this knowledge-constitutive interest, indeed of critical theory itself, is to restore to consciousness those suppressed, repressed and submerged determinants of unfree behaviour with a view to their dissolution (Habermas, 1984, pp. 194–5). This is a transformative agenda, concerned to move from oppression and inequality in society to the bringing about of social justice, equity and equality. These concern fairness in the egalitarian distribution of opportunities for, uptake of, processes in, participation in and outcomes of education and its impact on society, together with distributive justice, social justice and equality.

Mertens (2007, p. 213) argues that a transformative paradigm enters into every stage of the research process, because it concerns an interrogation of power. A transformative paradigm, she avers (pp. 216, 224), has several 'basic beliefs':

- Ontology (the nature of reality or of a phenomenon): politics and interests shape multiple beliefs and values, as these beliefs and values are socially constructed, privileging some views of reality and under-representing others;
- Epistemology (how we come to know these multiple realities): influenced by communities of practice which define what counts as acceptable ways of knowing, and affecting the relationships between the researcher and the communities who are being researched, such that partnerships are formed that are based on equality of power and esteem;
- Methodology (how we research complex, multiple realities): influenced by communities of practice which define what counts as acceptable ways of researching, and in which mixed methods can feature, as they enable a qualitative dialogue to be established between the participants in the research;
- Axiology (principles and meanings in conducting research, and the ethics that govern these): beneficence, respect and the promotion of social justice (see Chapter 7).

Mertens (p. 220) argues for mixed methods in a transformative paradigm (discussed later), as they reduce the privileging of powerful voices in society, and she suggests that participatory action research is a necessary, if not sufficient, element of a transformative paradigm, as it involves people as equals.

From Habermas's early work we conceptualize three research styles: the scientific, positivist style; the interpretive style; and the emancipatory, ideology critical style. Not only does critical theory have its own research agenda, but it also has its own research methodologies, in particular ideology critique and action research. The three methodologies, then, aligned to Habermas's knowledge-constitutive interests, are set out in Table 3.1.

With regard to ideology critique, a particular reading of ideology is being adopted here, as the 'suppression of generalizable interests' (Habermas, 1976, p. 113), where systems, groups and individuals operate in

TABLE 3.1 HABERMAS'S KNOWLEDGE-CONSTITUTIVE INTERESTS AND THE NATURE OF RESEARCH

Interest	*Methodology*	*Characteristics*
Technical interest	Scientific testing and proof	Scientific methodology; positivist (e.g. surveys, experiments); hypothesis testing; quantitative.
Practical interest	Hermeneutic; interpretive, understanding	Interactionist; phenomenological; humanistic; ethnographic; existential; anthropological; naturalistic; narratives; qualitative.
Emancipatory interest	Ideology critique	Political agenda, interrogation of power, transformative potential: people gaining control over their own lives; concern for social justice and freedom from oppression and from the suppression of generalizable interests; research to change society and to promote democracy.

rationally indefensible ways because their power to act relies on the disempowering of other groups, i.e. their principles of behaviour cannot be generalized.

Ideology – the values and practices emanating from particular dominant groups – is the means by which powerful groups promote and legitimate their particular – sectoral – interests at the expense of disempowered groups. Ideology critique exposes the operation of ideology in many spheres of education, the working out of vested interests under the mantle of the general good. The task of ideology critique is to uncover the vested interests at work that may be occurring consciously or subliminally, revealing to participants how they may be acting to perpetuate a system which keeps them either empowered or disempowered (Geuss, 1981), i.e. which suppresses a 'generalizable interest'. Explanations for situations might be other than those 'natural', taken for granted, explanations that the participants might offer or accept. Situations are not natural but problematic (Carr and Kemmis, 1986). They are the outcomes or processes wherein interests and powers are protected and suppressed; one task of ideology critique is to expose this (Grundy, 1987). The interests at work are uncovered by ideology critique, which, itself, is premised on reflective practice (Morrison, 1995a, 1995b, 1996a). Habermas (1972, p. 230) suggests that ideology critique through reflective practice can be addressed in four stages:

Stage 1: a description and interpretation of the existing situation – a hermeneutic exercise that identifies and attempts to make sense of the current situation (echoing the *verstehen* approaches of the interpretive paradigm).

Stage 2: a presentation of the reasons that brought the existing situation to the form that it takes – the causes and purposes of a situation and an evaluation of their legitimacy, involving an analysis of interests and ideologies at work in a situation, their power and legitimacy (both in micro- and macro-sociological terms). Habermas's early work (1972) likens this to psychoanalysis as a means for bringing into the consciousness of 'patients' those repressed, distorted and oppressive conditions, experiences and factors that have prevented them from having a full, complete and accurate understanding of their conditions, situations and behaviour, and that, on such exposure and examination, will be liberating and emancipatory. Critique here reveals to individuals and groups how their views and practices might be ideological distortions that, in their effects, perpetuate a social order or situation that works against their democratic freedoms, interests and empowerment (see also Carr and Kemmis, 1986, pp. 138–9).

Stage 3: an agenda for altering the situation – in order for moves to an egalitarian society to be furthered (the 'transformative paradigm' mentioned earlier).

Stage 4: an evaluation of the achievement of the situation in practice.

In the world of education, Habermas's stages are paralleled by Smyth (1989), who also denotes a four-stage process: *description* (what am I doing?); *information* (what does it mean?); *confrontation* (how did I come to be like this?); and *reconstruction* (how might I do things differently?). Ideology critique here has both a reflective, theoretical side and a practical side to it; without reflection it is blind and without practice it is empty.

As ideology is not mere theory but impacts directly on practice (Eagleton, 1991), there is a strongly practical methodology implied by critical theory, which articulates with action research (Callawaert, 1999). Action research (see Chapter 22), as its name suggests, is about research that impacts on, and focuses on, practice. In its espousal of practitioner research, for example, teachers in schools, participant observers and curriculum developers, action research recognizes the significance of *contexts* for practice – locational, ideological, historical, managerial, social. Further, it accords power to those who are operating in those contexts, for they are both the engines of research and of practice. The claim is made that action research is strongly empowering and emancipatory in that it gives practitioners a 'voice' (Carr and Kemmis, 1986; Grundy, 1987), participation in decision making and control over their environment and professional lives. Whether the strength of the claims for empowerment are as strong as their proponents would hold is another matter, for action research might be relatively powerless in the face of mandated changes in education and might be more concerned with intervening in existing practice to ensure that mandated change is addressed efficiently and effectively.

3.2 Criticisms of approaches from critical theory

Morrison (1995a) suggests that critical theory, because it has a practical intent to transform and empower, can – and should – be examined and perhaps tested empirically. For example, critical theory claims to be empowering; that is a testable proposition. Indeed, in a departure from some of his earlier writing, Habermas (1990) acknowledges this, arguing for the need to find 'counter examples' (p. 6), for 'critical testing' (p. 7) and empirical verification (p. 117). He acknowledges that his views have only 'hypothetical status' (p. 32)

that need to be checked against specific cases (p. 9). One could suggest, for instance, that the effectiveness of his critical theory can be examined by charting the extent to which: (a) equality, freedom, democracy, emancipation, empowerment have been realized by his theory; (b) transformative practices have been addressed or occurred as a result of his theory; (c) subscribers to his theory have been able to assert their agency; and (d) his theories have broken down the barriers of instrumental rationality. The operationalization and testing (or empirical investigation) of his theories clearly is a major undertaking. Without this, critical theory, a theory that strives to improve practical living, runs the risk of becoming merely contemplative.

There are several criticisms that have been voiced against critical approaches. Morrison (1995a) suggests that there is an artificial separation between Habermas's three interests – they are drawn far too sharply (Hesse, 1982; Bernstein, 1983, p. 33). For example, one has to bring hermeneutic knowledge to bear on positivist science and vice versa in order to make meaning of each other and in order to judge their own status. Further, the link between ideology critique and emancipation is neither clear nor proven, nor a logical necessity (Morrison, 1995a, p. 67) – whether a person or society can become emancipated simply by the exercise of ideology critique or action research is an empirical rather than a logical matter (Morrison, 1995a; Wardekker and Miedama, 1997). Indeed one can become emancipated by means other than ideology critique; emancipated societies do not necessarily demonstrate or require an awareness of ideology critique. Moreover, it could be argued that the rationalistic appeal of ideology critique actually obstructs action designed to bring about emancipation. Roderick (1986, p. 65), for example, questions whether the espousal of ideology critique is itself as ideological as the approaches that it proscribes. Habermas, in his allegiance to the social construction of knowledge through 'interests', is inviting the charge of relativism.

Whilst the claim to there being three forms of knowledge has the epistemological attraction of simplicity, one has to question this very simplicity (e.g. Keat, 1981, p. 67); there are a multitude of interests and ways of understanding the world and it is simply artificial to reduce these to three. Indeed it is unclear whether Habermas, in his three knowledge-constitutive interests, is dealing with a conceptual model, a political analysis, a set of generalities, a set of transhistorical principles, a set of temporally specific observations, or a set of loosely defined slogans (Morrison, 1995a, p. 71) that survive only by dint of their ambiguity (Kolakowsi, 1978). Lakomski (1999) questions the acceptability of the consensus theory of truth on which Habermas's work is premised (pp. 179–82); she argues that Habermas's work is silent on social change, and is little more than speculation and idealism, a view echoed by Fendler's (1999) criticism of critical theory as inadequately problematizing subjectivity and ahistoricity.

More fundamental to a critique of this approach is the view that critical theory has a deliberate political agenda, and that the task of the researcher is not to be an ideologue or to have an agenda, but to be dispassionate, disinterested and objective (Morrison, 1995a). Of course, critical theorists would argue that the call for researchers to be ideologically neutral is itself ideologically saturated with laissez-faire values which allow the status quo to be reproduced, i.e. that the call for researchers to be neutral and disinterested is just as value-laden as is the call for them to intrude their own perspectives. The rights of the researcher to move beyond disinterestedness are clearly contentious, though the safeguard here is that the researcher's is only one voice in the community of scholars (Kemmis, 1982). Critical theorists as researchers have been hoisted by their own petard, for if they are to become more than merely negative Jeremiahs and sceptics, berating a particular social order that is dominated by scientism and instrumental rationality (Eagleton, 1991; Wardekker and Miedama, 1997), they have to generate a positive agenda, but in so doing they are violating the traditional objectivity of researchers. Because their focus is on an ideological agenda, they themselves cannot avoid acting ideologically (Morrison, 1995a).

Claims have been made for the power of action research to empower participants as researchers (e.g. Carr and Kemmis, 1986; Grundy, 1987). This might be over-optimistic in a world in which power often operates through statute; the reality of political power seldom extends to teachers. That teachers might be able to exercise some power in schools but with little effect on society at large was caught in Bernstein's famous comment (1970) that 'education cannot compensate for society'. Giving action researchers a small degree of power (to research their own situations) has little effect on the *real* locus of power and decision making, which often lies outside the control of action researchers. Is action research genuinely and full-bloodedly empowering and emancipatory? Where is the evidence?

3.3 Participatory research and critical theory

The call to action in research, particularly in terms of participatory action by and with oppressed, disempowered, underprivileged and exploited groups, finds its

research voice in terms of participatory research (PR) (e.g. Freire, 1972; Giroux, 1989). Here the groups (e.g. community groups) themselves establish and implement interventions to bring about change, development and improvement to their lives, acting collectively rather than individually.

PR, an instance of critical theory in research, breaks with conventional ways of construing research, as it concerns doing research *with* people and communities rather than doing research *to* or *for* people and communities. It is premised on the view that research can be conducted by everyday people rather than an elite group of researchers, and that ordinary people are entirely capable of reflective and critical analysis of their situation (Pinto, 2000, p. 7). It is profoundly democratic, with all participants as equals; it strives for a participatory rather than a representative democracy (Giroux, 1983, 1989). PR regards power as shared and equalized, rather than as the property of an elite, and the researcher shares his or her humanity with the participants (Tandon, 2005a, p. 23). In PR, the emphasis is on research for change and development of communities; emphasis is placed on knowledge that is useful in improving lives rather than for the interests of, and under the control of, the academic or the researcher. It is research with a practical intent, for transforming lives and communities; it makes the practical more political and the political more practical (cf. Giroux, 1983). As Tandon (2005a, p. 23) writes: 'the very act of inquiry tends to have some impact on the social system under study'.

Campbell (2002, p. 20) suggests that PR arose as a reaction to those researchers and developers who adopted a 'top-down' approach to working with local communities, neglecting and relegating their local knowledge and neglecting their empowerment and improvement. Rather, PR is emancipatory (p. 20), eclectic and, like mixed methods research, adopts whatever research methodology will deliver the results that enable action and local development to follow. As with mixed methods research and action research, it is pragmatic, and, if necessary, sacrifices 'rigorous control, for the sake of "pragmatic utility" ' (Brown, 2005a, p. 92). PR challenges the conventional distance between researchers and participants; *together* they work for local development. It focuses on micro-development rather than macro-development, using knowledge to pursue well-being (Tandon, 2005b, p. ix; Brown, 2005a, p. 98).

PR respects the indigenous, popular knowledge that resides in communities rather than the relatively antiseptic world and knowledge of the expert researcher. Like Freire's work it is itself educative. Local community knowledge is legitimized in PR (Pinto, 2000, p. 21), and participants are active and powerful in the research rather than passive subjects. Local people can transform their lives through knowledge and their use of that knowledge; knowledge is power, with local community members collectively being active and in control. Researchers are facilitators, catalysts and change agents rather than assuming dominatory or controlling positions (Pinto, 2000, p. 13). The agenda of PR is empowerment of all and liberation from oppression, exploitation and poverty. Research here promotes both understanding and change. As one of its proponents, Lewin (1946, p. 34), wrote: 'if you want truly to understand something, try to change it'. PR blends knowledge and action (Tandon, 2005c, p. 49).

PR recognizes the centrality of power in research and everyday life, and has an explicit agenda of wresting power from those elites who hold it, and returning it to the grass roots, the communities, the mainstream citizenry. As Pinto (2000, p. 13) remarks, a core feature that runs right through all stages of PR is the nagging question of 'who controls?'.

PR has as its object the betterment of communities, societies and groups, often the disempowered, oppressed, impoverished and exploited communities, groups and societies, the poor, the 'have-nots' (Hall, 2005, p. 10; Tandon, 2005c, p. 50). Its principles concern improvement, group decision making, the need for research to have a practical outcome that benefits communities and in which participants are agents of their own decisions (Hall, 2005, p. 10; INCITE, 2010). It starts with problems as experienced in the local communities or workplace, and brings together into an ongoing working relationship both researchers and participants. As Bryeson *et al*. (2005, p. 183) remark, PR is a 'three-pronged activity' in which the investigation has the full and active participation by the community in question, involves action for development and which is an 'educational process of mobilization for development', and in which these three elements are interwoven. These features enter all stages of the research, from identification of problems to the design and implementation of the research, data analysis, reporting and catalysed changes and developments in the community. Empowerment and development are both the medium and the outcome of the research. Tandon (2005c, p. 30) sets out a sequence for PR (Figure 3.1).

Whilst conventional approaches to data collection may have their value (e.g. surveys, interviews), too often these are instruments that regard people solely as sources of information rather than as participants in their own community development (Hall, 2005, p. 13). Indeed Tandon (2005d, p. 106) reports that, in many

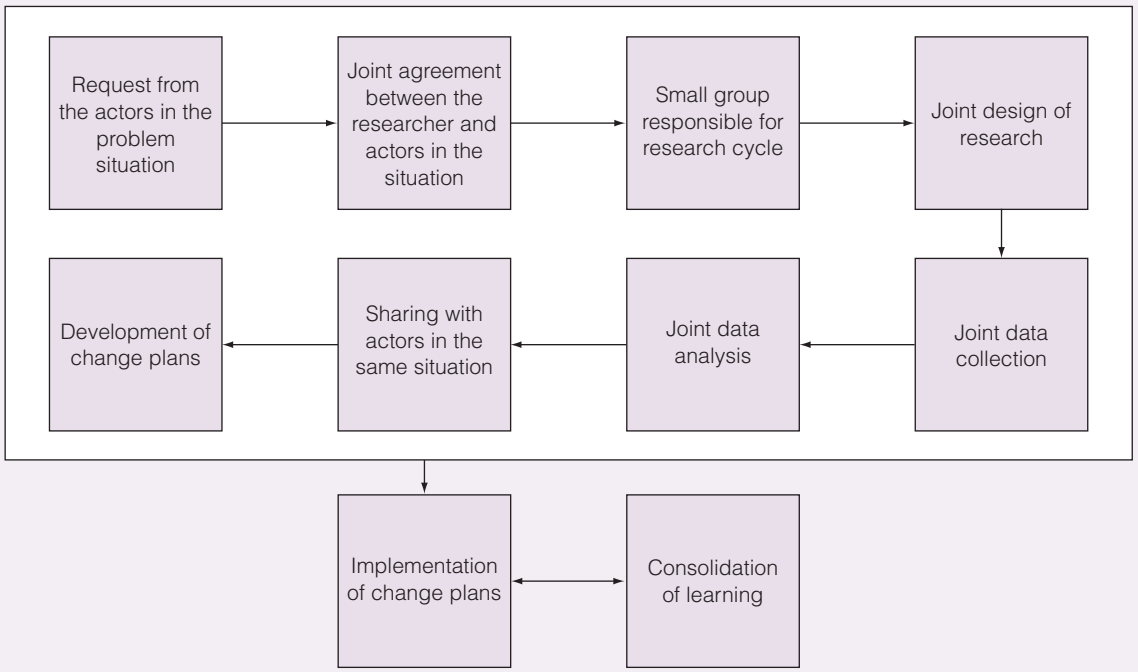


FIGURE 3.1 Steps in an 'ideal' participatory research approach

Source: Tandon (2005c, p. 30), reproduced with permission of Mosaic Books, New Delhi

cases, surveys are entirely irrelevant to the communities involved in the research, and alternative forms of collecting data have to be used, for example, dialogue (Tandon, 2005e), enumeration such as census data (though clearly these are used in conventional research) (Batliwala and Patel, 2005), and popular theatre for consciousness-raising (Khot, 2005). Hall (2005) cites the example of the UNESCO evaluation of the Experimental World Literacy Programme, in which local expertise was neglected, which over-simplified the phenomena under investigation and disempowered the very communities under review. Such research is alienating rather than empowering. Rather, he avers, researchers should respect, and take seriously, resident knowledge (he gives the example of adult learning).

Hall (2005, pp. 17–19) sets out several principles for PR:

1 A research project – both process and results – can be of immediate and direct benefit to a community (as opposed to serving merely as the basis of an academic paper of obscure policy analysis).
2 A research project should involve the community in the entire research project, from the formulation of the problem and the interpretation of the findings to planning corrective action based upon them.
3 The research process should be seen as part of a total educational experience which serves to determine community needs, and to increase awareness of problems and commitment to solutions within the community.
4 Research should be viewed as a dialectic process, a dialogue over time, and not a static picture of reality at one point in time.
5 The object of research, like the object of education, should be the liberation of human creative potential and the mobilization of human resources for the solution of social problems.
6 Research has ideological implications.… First is the re-affirmation of the political nature of all we do.… Research that allows for popular involvement and increased capacities of analysis will also make conflictual action possible, or necessary.

(Hall, 2005, pp. 17–19)

In PR the problem to be investigated originates in, and is defined by, the community or workplace. It members are involved in the research and have control over it,

and the research leads to development and improvement of their lives and communities (Brown and Tandon, 2005, p. 55). Brown and Tandon (p. 60) recognize the challenge (and likely resistance) that these principles might pose for the powerful, specific dominant interest groups, but they argue that this is unavoidable, as the researcher typically mobilizes community groups to action (p. 61). Hence PR has to consider the likely responses of the researchers, the participants and their possible opponents (p. 62); as Giroux avers (1983), knowledge is not only powerful, but dangerous, and participants may run substantial risks (Brown and Tandon, 2005, p. 65) in conducting this type of research, for it upsets existing power structures in society and the workplace.

PR has some affinity to action research (INCITE, 2010), though it is intensely more political than action research. It is not without its critics. For example, Brown (2005b) argues that participatory action research is ambiguous about:

a its research objectives (e.g. social change, raising awareness, development work, challenging conventional research paradigms);
b the relationships between the researcher and participants (e.g. over-emphasizing similarities and neglecting differences between them);
c the methods and technologies that it uses (e.g. being over-critical of conventional approaches which might serve the interests of participatory research, and the lack of a clear method for data collection); and
d the outcomes of participatory research (e.g. what these are, when these are decided, and who decides).

Notwithstanding these, however, PR is a clear instance of the tenets of critical theory, transformative action and empowerment put into practice.

3.4 Feminist research

It is no mere coincidence that feminist research should surface as a serious issue at the same time as ideology-critical paradigms for research; they are closely connected. Usher (1996) sets out several principles of feminist research that resonate with the ideology critique of the Frankfurt School:

- the acknowledgement of the pervasive influence of gender as a category of analysis and organization;
- the deconstruction of traditional commitments to truth, objectivity and neutrality;
- the adoption of an approach to knowledge creation which recognizes that all theories are perspectival;
- the utilization of a multiplicity of research methods;
- the inter-disciplinary nature of feminist research;
- involvement of the researcher and the people being researched;
- the deconstruction of the theory/practice relationship.

Her suggestions build on the recognition of the significance of addressing the 'power issue' in research ('whose research', 'research for whom', 'research in whose interests') and the need to address the emancipatory element of educational research: research should be empowering to all participants. Critical theory questions the putative objective, neutral, value-free, positivist, 'scientific' paradigm for the sundering of theory and practice and for its reproduction of asymmetries of power (reproducing power differentials in the research community and for treating participants/respondents instrumentally, as objects).

Robson (1993, p. 64) suggests seven sources of sexism in research:

- *androcentricity*: seeing the world through male eyes and applying male research paradigms to females;
- *overgeneralization*: when a study generalizes from males to females;
- *gender insensitivity*: ignoring gender as a possible variable;
- *double standards*: using male criteria, measures and standards to judge the behaviour of women and vice versa (e.g. in terms of social status);
- *sex appropriateness*: for example, that child-rearing is women's responsibility;
- *familism*: treating the family, rather than the individual, as the unit of analysis;
- *sexual dichotomism*: treating the sexes as distinct social groups when, in fact, they may share characteristics.

Feminist research challenges the legitimacy of research that does not empower oppressed and otherwise invisible groups – women. Ezzy (2002, p. 20) writes of the need to replace a traditional masculine picture of science with an emancipatory commitment to knowledge that stems from a feminist perspective, since, if researchers analyse women's experiences 'using only theories and observations from the standpoint of men, the resulting theories oppress women' (p. 23). Gender, as Ezzy writes (p. 43), is 'a category of experience'.

Positivist research serves a given set of power relations, typically empowering the white, male-dominated research community at the expense of other groups whose voices are silenced. Feminist research seeks to demolish and replace this with a different substantive

agenda of empowerment, voice, emancipation, equality and representation for oppressed groups. In doing so, it recognizes the necessity for foregrounding issues of power, silencing and voicing, ideology critique and a questioning of the legitimacy of research that does not emancipate hitherto disempowered groups. In feminist research, women's consciousness of oppression, exploitation and disempowerment becomes a focus for research and ideology critique.

Far from treating educational research as objective and value-free, feminists argue that this is merely a smokescreen that serves the existing, disempowering status quo, and that the subject and value-laden nature of research must be surfaced, exposed and engaged (Haig, 1999, p. 223). Supposedly value-free, neutral research perpetuates power differentials. Indeed Jayaratne and Stewart (1991) question the traditional, exploitative nature of much research in which the researchers receive all the rewards whilst the participants remain in their – typically powerless – situation, i.e. in which the status quo of oppression, underprivilege and inequality remain undisturbed. Scott (1985, p. 80) writes that 'we may simply use other women's experiences to further our own aims and careers' and questions how ethical it is for a woman researcher to interview those who are less privileged and more exploited than she herself is. Creswell (1998, p. 83), too, suggests that feminist research strives to establish collaborative and non-exploitative relationships.

Researchers, then, must take seriously issues of reflexivity, the effects of the research on the researched and the researchers, the breakdown of the positivist paradigm, and the raising of consciousness of the purposes and effects of the research. Ezzy (2002, p. 153) notes that an integral element of the research is the personal experience of the researcher himself/herself, reinforcing the point that objectivity is a false claim by researchers.

Denzin (1989), Mies (1993), Haig (1999) and De Laine (2000) argue for several principles in feminist research:

- the asymmetry of gender relations and representation must be studied reflexively as constituting a fundamental aspect of social life (which includes educational research);
- women's issues, their history, biography and biology, feature as a substantive agenda/focus in research – moving beyond mere perspectival/methodological issues to setting a research agenda;
- the raising of consciousness of oppression, exploitation, empowerment, equality, voice and representation is a methodological tool;
- the acceptability and notion of objectivity and objective research must be challenged;
- the substantive, value-laden dimensions and purposes of feminist research must be paramount;
- research must empower women;
- research need not only be undertaken by academic experts;
- collective research is necessary – women need to collectivize their own individual histories if they are to appropriate these histories for emancipation;
- there is a commitment to revealing core processes and recurring features of women's oppression;
- an insistence on the inseparability of theory and practice;
- an insistence on the connections between the private and the public, between the domestic and the political;
- a concern with the construction and reproduction of gender and sexual differences;
- a rejection of narrow disciplinary boundaries;
- a rejection of the artificial subject/researcher dualism;
- a rejection of positivism and objectivity as male mythology;
- the increased use of qualitative, introspective biographical research techniques;
- a recognition of the gendered nature of social research and the development of anti-sexist research strategies;
- a review of the research process as consciousness and awareness raising and as fundamentally participatory;
- the primacy of women's personal subjective experience;
- the rejection of hierarchies in social research;
- the vertical, hierarchical relationships of researchers/research community and research objects, in which the research itself can become an instrument of domination and the reproduction and legitimation of power elites, must be replaced by research that promotes the interests of dominated, oppressed, exploited groups;
- the recognition of equal status and reciprocal relationships between subjects and researchers;
- the need to change the status quo, not merely to understand or interpret it;
- the research must be a process of conscientization, not research solely by experts for experts, but to empower oppressed participants.

Webb *et al*. (2004) set out six principles for a feminist pedagogy in the teaching of research methodology:

1 reformulation of the professor–student relationship (from hierarchy to equality and sharing);

2 empowerment (for a participatory democracy);
3 building community (through collaborative learning);
4 privileging the individual voice (not only the lecturer's);
5 respect for diversity of personal experience (rooted, for example, in gender, race, ethnicity, class, sexual preference);
6 challenging traditional views (e.g. the sociology of knowledge).

Gender shapes research agendas, the choice of topics, foci, data-collection techniques and the relationships between researchers and researched. Several methodological principles flow from a 'rationale' for feminist research (Denzin, 1989; Mies, 1993; Haig, 1997, 1999; De Laine, 2000):

- the replacement of quantitative, positivist, objective research with qualitative, interpretive, ethnographic reflexive research, as objectivity in quantitative research is a smokescreen for masculine interests and agendas;
- collaborative, collectivist research undertaken by collectives – often of women – combining researchers and researched in order to break subject/object and hierarchical, non-reciprocal relationships;
- the appeal to alleged value-free, neutral, indifferent and impartial research has to be replaced by conscious, deliberate partiality – through researchers identifying with participants;
- the use of ideology-critical approaches and paradigms for research;
- the spectator theory or contemplative theory of knowledge in which researchers research from ivory towers must be replaced by a participatory approach – for example, action research – in which all participants (including researchers) engage in the struggle for women's emancipation – a liberatory methodology;
- the need to change the status quo is the starting point for social research;
- the extended use of triangulation, multiple methods (including visual techniques such as video, photography and film), linguistic techniques such as conversational analysis and of textual analysis such as deconstruction of documents and texts about women;
- the use of meta-analysis to synthesize findings from individual studies (see Chapter 21);
- a move away from numerical surveys and a critical evaluation of them, including a critique of question wording.

Edwards and Mauthner (2002, pp. 15, 27) characterize feminist research as that which concerns a critique of dominatory and value-free research, the surfacing and rejection of exploitative power hierarchies between the researcher and the participants, and the espousal of close – even intimate – relationships between the researcher and the researched. Positivist research is rejected as per se oppressive (Gillies and Alldred, 2002, p. 34) and inherently unable to abide by its own principle of objectivity; it is a flawed epistemology. Research and its underpinning epistemologies are rooted in, and inseparable from, interests (Habermas, 1972).

The move is towards 'participatory action research' which promotes empowerment and emancipation and which is an involved, engaged and collaborative process (e.g. De Laine, 2000, pp. 109ff.). Participation recognizes imbalances of power and the imperative to 'engage oppressed people as agents of their own change' (Ezzy, 2002, p. 44), whilst action research recognizes the value of utilizing the findings from research to inform decisions about interventions (p. 44). As De Laine (2000, p. 16) writes, the call is for 'more participation and less observation, of *being with* and *for* the other, not *looking at*', with relations of reciprocity and equality rather than impersonality, exploitation and power/status differentials between researcher and participants.

The relationship between the researcher and participant, De Laine argues, must break a conventional patriarchy. The emphasis is on partnerships between researchers and participants (p. 107), with researchers as participants rather than outsiders and with participants shaping the research process as co-researchers (p. 107), defining the problem, methods, data collection and analysis, interpretation and dissemination. The relationship between researchers and participants is one of equality, and outsider, objective, distant, positivist research relations are off the agenda; researchers are inextricably bound up in the lives of those they research. That this may bring difficulties in participant and researcher reactivity is a matter to be engaged rather than built out of the research.

Thapar-Björkert and Henry (2004) argue that the conventional, one-sided and unidirectional view of the researcher as powerful and the research participants as less powerful, with the researcher exploiting and manipulating the researched, could be a construction by western white researchers. They report research that indicates that power is exercised by the researched as well as the researchers, and is a much more fluid, shifting and negotiated matter than conventionally suggested, being dispersed through both the researcher and

the researched. Indeed they show how the research participants can, and do, exercise considerable power over the researchers, both before, during and after the research process. They provide a fascinating example of interviewing women in their homes in India, where, far from being a location of oppression, the home was a site of their power and control.

With regard to methods of data collection, Oakley (1981) suggests that interviewing women in the standardized, impersonal style which expects a response to a prescribed agenda and set of questions may be a 'contradiction in terms', as it implies an exploitative relationship. Rather, the subject/object relationship should be replaced by a guided dialogue. She criticizes the conventional notion of 'rapport' in conducting interviews (p. 35), arguing that such interviews are instrumental, non-reciprocal and hierarchical, all of which are masculine traits. Rapport in this sense, she argues, is not genuine in that the researcher is using it for scientific rather than human ends (p. 55). Here researchers are 'faking friendship' for their own ends (Duncombe and Jessop, 2002, p. 108), equating 'doing rapport' with trust, and, thereby, operating a very 'detached' form of friendship (p. 110) (see also Thapar-Björkert and Henry, 2004).

Duncombe and Jessop (2002, p. 111) question whether, if interviewees are persuaded to take part in an interview by virtue of the researcher's demonstration of empathy and 'rapport', this is really giving informed consent. They suggest that informed consent, particularly in exploratory interviews, has to be continually renegotiated and care has to be taken by the interviewer not to be too intrusive. Personal testimonies, oral narratives and long interviews also figure highly in feminist approaches (De Laine, 2000, p. 110; Thapar-Björkert and Henry, 2004), not least in those which touch on sensitive issues. These, it is argued (Ezzy, 2002, p. 45), enable women's voices to be heard, to be close to lived experiences, and avoid unwarranted assumptions about people's experiences.

The drive towards collective, egalitarian and emancipatory qualitative research is seen as necessary if women are to avoid colluding in their own oppression by undertaking positivist, uninvolved, dispassionate, objective research. Mies (1993, p. 67) argues that for women to undertake this latter form of research puts them into a schizophrenic position of having to adopt methods which contribute to their own subjugation and repression by ignoring their experience (however vicarious) of oppression and by forcing them to abide by the 'rules of the game' of the competitive, male-dominated academic world. In this view, argue Roman and Apple (1990, p. 59), it is not enough for women simply to embrace ethnographic forms of research, as this does not necessarily challenge the existing and constituting forces of oppression or asymmetries of power. Ethnographic research, they argue, has to be accompanied by ideology critique; indeed they argue that the transformative, empowering, emancipatory potential of research is a critical standard for evaluating the research.

This latter point resonates with the call by Lather (1991) for researchers to be concerned with the political consequences of their research (e.g. consequential validity), not only the conduct of the research and data analysis itself. Research must lead to change and improvement for women (Gillies and Alldred, 2002, p. 32). Research is a political activity with a political agenda (p. 33; see also Lather, 1991). Research and action – praxis – must combine: 'knowledge *for*' as well as 'knowledge *what*' (Ezzy, 2002, p. 47). As Marx reminds us in his *Theses on Feuerbach*: 'the philosophers have only interpreted the world, in various ways; the point, however, is to change it'. Gillies and Alldred (2002, p. 45), however, point out that 'many feminists have agonized over whether politicizing participants is necessarily helpful', as it raises awareness of constraints on their actions without being able to offer solutions or to challenge their structural causes. Research, thus politicized but unable to change conditions, may actually be disempowering and, indeed, patronizing in its simplistic call for enlightenment and emancipation. It could render women more vulnerable than before. Emancipation is a struggle.

Several of these views of feminist research and methodology are contested by other feminist researchers. For example, Jayaratne (1993, p. 109) argues for 'fitness for purpose', suggesting that an exclusive focus on qualitative methodologies might not be appropriate either for the research purposes or, indeed, for advancing the feminist agenda (see also Scott, 1985, pp. 82–3). Jayaratne refutes the argument that quantitative methods are unsuitable for feminists because they neglect the emotions of the people under study. Indeed she argues for beating quantitative research on its own grounds (1993, p. 121), suggesting the need for feminist quantitative data and methodologies in order to counter sexist quantitative data in the social sciences. She suggests that feminist researchers can accomplish this without 'selling out' to the positivist, male-dominated academic research community. Indeed Oakley (1998) suggests that the separation of women from quantitative methodology may have the unintended effect of perpetuating women as the 'other', and, thereby, discriminating against them. Finch (2004) argues that, whilst qualitative research might have

helped to establish the early feminist movement, it is important to recognize the place of both quantitative and qualitative methods as the stuff of feminist research.

De Laine (2000, p. 1132) reports work that suggests that close relationships between researchers and participants may be construed as being as exploitative, if disguised, as conventional researcher roles, and that they may bring considerable problems if data that were revealed in an intimate account between friends (researcher and participant) are then used in public research. The researcher is caught in a dilemma: if she is a true friend then this imposes constraints on the researcher, and yet if she is only pretending to be a friend, or limiting that friendship, then this provokes questions of honesty and personal integrity. Are research friendships real, ephemeral or impression management used to gather data?

De Laine (p. 115) suggests that it may be misguided to privilege qualitative research for its claim to non-exploitative relationships. Whilst she acknowledges that quantitative approaches may perpetuate power differentials and exploitation, there is no guarantee that qualitative research will not do the same, only in a more disguised way. Qualitative approaches too, she suggests, can create and perpetuate unequal relations, not least simply because the researcher is in the field qua researcher rather than a friend; if it were not for the research then the researcher would not be present. Stacey (1988) suggests that the intimacy advocated for feminist ethnography may render exploitative relationships *more* rather than *less* likely. We refer readers to Chapter 13 on sensitive educational research for a further discussion of these issues. Ezzy (2002, p. 44) reports that, just as there is no single feminist methodology, both quantitative and qualitative methods are entirely legitimate. Indeed Kelly (1978) argues that a feminist commitment should enter research at the stages of formulating the research topic and interpreting the results, but it should be left out during the stages of data collection and conduct of the research.

Gillies and Alldred (2002, pp. 43–6) suggest that action research, an area strongly supported by some feminist researchers, is itself problematic. It risks being an intervention in people's lives (i.e. a potential abuse of power), and the researcher typically plays a significant, if not central, role in initiating, facilitating, crystallizing and developing the meanings involved in, or stemming from, the research, i.e. the researcher is the one exercising power and influence.

Thapar-Björkert and Henry (2004) indicate that the researcher being an outsider might bring more advantages than if she were an insider. For example, being a white female researching non-white females may not be a handicap, as many non-white women might disclose information to white women that they would not disclose to a non-white person. Similarly, having interviewers and interviewees of the same racial and ethnic background does not mean that non-hierarchical relationships will not still be present. They also report that the categories of 'insider' and 'outsider' are much more fuzzy than exclusive. Researchers are both 'subject' and 'object', and those being researched are both 'observed' and 'observers'.

De Laine (2000, p. 110) suggests that there is a division among feminists between those who advocate closeness in relationships between researchers and subjects – a human researching fellow humans – and those who advocate 'respectful distance' between researchers and those being studied. Duncombe and Jessop (2002, p. 111) comment that close relationships may turn into quasi-therapeutic situations rather than research, yet it may be important to establish closeness in reaching deeper issues, and they question how far close relationships lead to reciprocal and mutual disclosure (p. 120). The debate is open: should the researcher share, be close and be prepared for more intimate social relations – a 'feminist ethic of care' (p. 111) – or keep those cool, outsider relations which might objectify those being researched? It is a moral as well as a methodological matter.

The issue runs deep: the suggestion is that emotions and feelings are integral to the research, rather than to be built out of the research in the interests of objectivity (Edwards and Mauthner, 2002, p. 19). Emotions should not be seen as disruptive of research or as irrelevant (De Laine, 2000, pp. 151–2), but central to it, just as they are central to human life; indeed emotional responses are essential in establishing the veracity of inquiries and data, and the 'feminist communitarian model' which De Laine outlines (pp. 212–13) values connectedness at several levels: emotions, emotionality and personal expressiveness, empathy. The egalitarian feminism that De Laine (2000) and others advocate suggests a community of insiders in the same culture, in which empathy, reciprocity and egalitarianism are hallmarks (p. 108).

Swantz (1996, p. 134) argues that there may be some self-deception by the researcher in adopting a dual role as a researcher and one who shares the situation and interests of the participants. She questions the extent to which the researcher may be able to be genuinely involved with the participants in other than a peripheral way and whether, simply because the researcher may have 'superior knowledge', a covert power differential may exist. De Laine (2000, p. 114)

suggests that such superior knowledge may stem from the researcher's own background in anthropology or ethnography, or simply more education. The primary purpose of the researcher is research, and that is different from the primary purpose of the participants.

The researcher's desire for identification and solidarity with her research subjects may be pious but unrealistic optimism, not least because she may not share the same race, ethnicity, background, life chances, experiences or colour as those being researched. Indeed Gillies and Alldred (2002, pp. 39–40) raise the question of how far researchers can, or should, try to represent groups to which they themselves do not belong, including those groups without power or voice, as this itself is a form of colonization and oppression. Affinity, they argue (p. 40), is no authoritative basis for representative research. Even the notion of affinity becomes suspect when it overlooks or underplays the significance of difference, thereby homogenizing groups and their particular experiences. In response to this, some feminist researchers (p. 40) suggest that researchers only have the warrant to confine themselves to their own immediate communities, though this is a contentious issue. There is value in speaking for others, not least for those who are silenced and marginalized, and in not speaking for others for fear of oppression and colonization. They also question the acceptability and appropriateness of, and fidelity to, the feminist ethic, if one represents and uses others' stories (p. 41).

3.5 A note on post-colonial theory and queer theory

Under the umbrella of critical theory also fall post-colonial theory, queer theory and critical race theory. Whilst this chapter does not unpack these, it notes them as avenues for educational researchers to explore. For example, post-colonial theory, as its name suggests, with an affinity to postmodernism, addresses the experiences (often through film, literature, cultural studies, political and social sciences) of post-colonial societies and the cultural legacies of colonialism. It examines the after-effects, or continuation, of ideologies and discourses of imperialism, domination and repression, value systems (e.g. the domination of western values and the delegitimization of non-western values), their effects on the daily lived experiences of participants, i.e. their materiality, and the regard in which peoples in post-colonial societies are held (e.g. Said's (1978) work on orientalism and the casting down of non-western groups as the 'other'). It also discusses the valorization of multiple voices and heterogeneity in post-colonial societies, the resistance to marginalization of groups within them (Bhabha, 1994, p. 113) and the construction of identities in a post-colonial world.

Queer theory builds on, but moves beyond, feminist theory and gay/lesbian/LGBTI studies to explore the social construction and privileging or denial of identities, sexual behaviour, deviant behaviour and the categorizations and ideologies involved in such constructions. It deconstructs 'social categories and binary identities' (Marshall and Rossman, 2016, p. 26) in striving to demonstrate that such categories are, in reality, more fluid and transparent than is often assumed or bounded. Identity, for queer theorists, is not singular, fixed and firm, but multiple, unstable and fluid, and that when applied to commonly held categories such as heterosexuality, it reveals such fluidity.

Halperin (1997) writes that queer theory focuses on whatever is 'at odds with the normal, the legitimate, the dominant' (p. 62). Its task is to explore, problematize and interrogate gender, sexual orientation and also their mediation by, and intersection with, other characteristics or forms of oppression, for example, social class, ethnicity, colour, disability, nationality, age, able-ness (Marshall and Rossman, 2016, p. 27). However, it does not confine itself to matters of sexuality but makes 'queer' a range of commonly held categories. It rejects simplistic categorizations of individuals, and argues for the respect of their individuality and uniqueness. Queer theory does not adhere to a single research method but advocates multiple methods which promote collaborative understandings and reflexivity on the part of research participants and researchers.

3.6 Value-neutrality in educational research

Lather (1986a) argues that, as neither education nor research is neutral, researchers do not need to apologize for undertaking clearly ideological research and its intention to change the status quo of inequality (p. 67). However, the case is made that research should be disinterested and objective, that value-neutrality is an ideal and that research should concern itself only with the pursuit and production of facts and knowledge and not play politics, but that this does not preclude value-relevant research, i.e. topics that may be of concern to certain parties. Politics and research are not the same and it is illegitimate for the researcher to let a political agenda enter into – to bias – the conduct of, and conclusions from, research.

However, it is argued that developments in the philosophy of science indicate that researchers make all kinds of assumptions about the world, both factual and

evaluative, and that these shape the research (Hammersley, 2000, p. 3), i.e. that there is no such thing as objective knowledge but only knowledge that is socio-culturally situated. This is the argument brought forward by post-positivism, postmodernism and post-structuralism, though Hammersley notes that, whilst values might, indeed maybe should, determine what is considered to be value-relevant research (i.e. what topics to focus on), and that this is completely within the scope of factual research, nevertheless 'research must necessarily be committed to value neutrality simply because it cannot validate value conclusions' (p. 32).

Should researchers be objective, value-neutral, non-partisan, unbiased and strictly disinterested, simply providing a service in bringing forward factual evidence, data and explanations on such-and-such a matter, or is it acceptable for them to declare their values, biases and interests and then proceed from there, acting on those commitments? Should researchers have a political or social agenda that colours their research? Should they be 'committed' or should they be disinterested? Hammersley's (2000) comments on 'standpoint epistemology' feature here (pp. 6–7), where he notes that, in Marxism, the working class is in a privileged position in understanding capitalist society and how it should be and can be transformed. Similarly he gives the example of women as oppressed or marginalized groups in patriarchal societies and he questions whether this might give them a position on and understanding of oppression and power that is simply not available to men (Hammersley, 2011, pp. 97–9).

Do we only ask white males about the experience of being a non-white woman, or do we only ask non-white women about their experiences, or do we ask both groups, since their perspectives and knowledge might differ? Is there any guarantee that any of these groups will see 'reality' clearly (cf. Hammersley, 2011, p. 99)? What warrant can be brought forward to justify the privileging of one group's views over those of another?

If a researcher happens to believe in democracy, social justice and equality, or free-market neo-liberalism, or communism, or is African-American or a white working-class female, should that affect how he or she conducts research and the conclusions and prescriptions that he or she draws from it? Should researchers push their own or others' political or social agendas?

Hammersley (2000) unpicked dangers of partisanship, 'committed' positions and 'privileged' discourse on the part of researchers as this can 'encourage the idea that research can, by itself, tell us what is desirable and undesirable, and what should be done; thereby obscuring the value judgements involved in policy and practice' (Hammersley, 2011 p. 87). He focuses on critical social science, particularly critical realism, noting that whilst value argument is important, indeed is essential to politics, social scientists 'have no distinctive authority to determine what is good or bad about the situations they seek to describe and explain; or what, if anything, should be done about them' (2014, p. 94). He argues (p. 94) that they, among other parties, have the authority of expertise concerning matters of fact but not to matters of value.

This echoes Weber's (1949) comment that an empirical science should not be committed to providing 'binding norms and ideals from which directives for immediate practical activity can be derived' (p. 52). Researchers may have their own political agendas or interests and these might determine their choice of areas of research, but that is an entirely different matter from saying that they will or should push their own views and personal political agendas, making prescriptions that emanate from their research for their own partisan agendas (cf. Pawson's (2013, pp. 61ff.) critique of critical realism for its disguised normative premises).

Phillips and Burbules (2000) note that whilst extra-scientific values might determine the focus of the research, this does not mean that those values should influence the conduct of the research (p. 53). Risjord (2014, p. 18) argues for 'epistemic values' (objective scientific reasoning) to be the hallmarks of research, and that these should not be confused with 'non-epistemic values' (moral and political values). Similarly Hammersley (2000, pp. 17–18) suggests that arguing against value-neutrality in research confuses the conduct of research (concerning itself with factual content) with its consequences and implications, and that, save for ethical limits, researchers do not have responsibility for what happens with regard to the consequences of their research. In other words, researchers remain disinterested and neutral, provide evidence, explanations and facts, even recommendations, but leave politics alone. Fact and value differ.

On the other hand, the question is raised that, by not addressing consequences and implications, researchers enable the status quo of inequality, social injustice and oppression to be perpetuated and that it is incumbent on researchers not to hide behind putative value-neutrality, because, in effect, such research is not value-neutral but reinforces the dominant ideology and the interests of the powerful (Hammersley, 2000, p. 136). One cannot pretend that oppression does not exist, and, therefore, to argue for value-neutrality demonstrates a political or moral commitment (Risjord, 2014, p. 28).

In response to this, however, the argument is brought forward that the nature of society is much more contested, complex, dissonant and unclear than critical theorists would argue, and, indeed, that their view of society is more an article of faith, an assumption or presupposition, a value or, indeed, a dogma or ideology that closes itself up to critical enquiry and sound knowledge, or that it harks back to the foundationalism so roundly criticized by post-positivists and post-structuralists. Social reality is not necessarily the taken-for-granted world as that seen through the eyes of critical theorists. In other words, critical theory may be as biased as those views of society it seeks to criticize, and to see society in such dichotomous, either/or terms – equal or unequal, socially just or socially unjust, democratic or undemocratic, free or unfree – or to see it as more complex but still characterized as being marked by oppression, ideology and injustice, is naive, not least as the same circumstances that gave rise to what critical theorists would call inequality also gave rise to greater equality. Just as there is no single, one-dimensional view of society and social reality, so there is no single view of how it must be viewed or researched. In this case, the researcher must regard the claims of critical theorists as hypotheses to be tested rather than as cases that are already proven.

Further, the terminology used by critical theorists is problematic (Hammersley, 2000, p. 139); terms such as 'equality', 'discrimination', 'inequality' are open to differences of interpretation, and, indeed, to differences in value. The same term has different meanings, interpretations and values; indeed, to derive values from facts is to conflate an 'is' with an 'ought', and this is not the stuff of research (see Hammersley's (2014) criticism of critical realism on these grounds).

Is the job of researchers only to provide evidence and explanation, or does it extend into promoting political agendas? Should researchers be partisan or non-partisan, 'committed' or 'disinterested'? Should their own political values or views of what society should be like enter into their research? Whilst objectivity and value-neutrality have been called into question by the post-positivists, indeed by many researchers, what is the limit of this? Here we have two distinct, perhaps irreconcilable views of the tasks and roles of the researcher and research: to provide information – to be a 'methodological purist' (Hammersley, 2000), or to be a political activist. Of course, serving political goals does not preclude the possibility that: (a) knowledge will be produced or facilitated by taking a political stance; (b) those who do not subscribe to the values or views of critical theorists are not simply 'ideological dopes of stunning mediocrity' (Giddens, 1979, p. 52); (c) those who are committed to value-neutrality are not free from the chance of making errors; (d) power differentials do exist in society regardless of which lens one uses to view it. Is there common ground between the analytical, value-neutral researcher and the partisan researcher, whether the latter espouses critical theory or some value system? Is it the case, as Hammersley (2000) so trenchantly puts it, that 'the critical approach disqualifies itself as a form of academic research: it turns sociology into a political morality play' (p. 150)?

3.7 A summary of three major paradigms

The three chapters so far have discussed very different approaches to educational research, which rest on quantitative, qualitative and critical theoretical foundations, or a combination of these.

Table 3.2 summarizes some of the broad differences between the approaches that we have made so far. We present the paradigms and their affiliates in Figure 3.2.

The companion website to the book provides PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: www.routledge.com/cw/cohen.

TABLE 3.2 DIFFERING APPROACHES TO THE STUDY OF BEHAVIOUR

Normative	*Interpretive*	*Complexity theoretical*	*Critical*
Society and the social system	The individual	Wholes, groups, systems and the individuals within them	Societies, groups and individuals
Medium/large-scale research	Small-scale research	Micro- and macro-scale research	Small-scale research
Impersonal, anonymous forces regulating behaviour	Human actions continuously recreating social life	Individuals and their environments constantly and dynamically interact to produce new, emergent systems and behaviours through self-organization, connectedness and feedback	Political, ideological factors, power and interests shaping behaviour
Model of natural sciences	Non-statistical	Action research, case study and narrative research	Ideology critique, action research and critical ethnography
Quantitative, objective	Qualitative, subjective	Quantitative, qualitative, objective and subjective, algorithmic	Ideology critique, participatory, objective and subjective
Positivist and scientific	Hermeneutic and interpretive	Systems-driven, social network driven	Ideology critical
Linear causality	Multiple directions of causality	Multiple directions of causality	Main trends of causality
Reductionist and atomistic	Phenomenologists, symbolic interactionists, ethnomethodologists	Holistic understanding of emergent conditions and systems	Change and emancipation
Research conducted 'from the outside'	Insider and outsider research	Non-reductionist	Interpretive, macro- and micro-concepts: political and ideological interests, operations of power
Outsider research	Personal involvement of the researcher	Objective analysis of systems	Critical theorists, action researchers, practitioner researchers
Generalizing from the specific	Interpreting the specific	Understanding wholes	Collectivity
Explaining behaviour/seeking causes	Understanding actions/meanings rather than causes	Understanding causal interactions	Participant researchers, researchers and facilitators
Assuming the taken-for-granted	Investigating the taken-for-granted	Investigating emergent systems	Critiquing the specific
Macro-concepts: society, institutions, norms, positions, roles, expectations	Micro-concepts: individual perspective, personal constructs, negotiated meanings, definitions of situations	Micro- and macro-level analysis informing each other	Understanding, interrogating, critiquing, transforming actions and interests
Structuralists	Hermeneutic	Explaining and observing, iterative	Interrogating, critiquing and changing the taken-for-granted
Prediction and control	Understanding and explanation	Understanding emergence of complex adaptive systems	Transformation and praxis
Technical interest	Practical interest	Technical and practical interest	Emancipatory interest

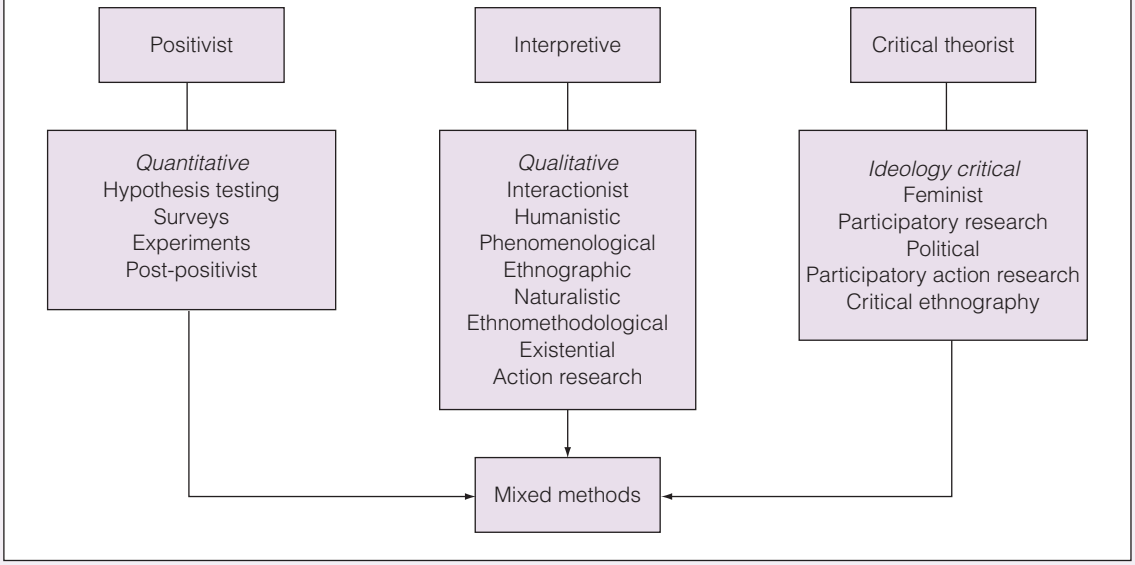


FIGURE 3.2 Positivist, interpretive and critical paradigms in educational research

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Theory in educational research

CHAPTER 4

Educational researchers are frequently exhorted to root their research in a theoretical framework. This short chapter explores what this means and addresses the following issues:

- What does 'theory' mean and how does it relate to research?
- Where does 'theory' come from, and how is it used? What is it for?
- Do we really need theory in conducting research?
- What kinds of theory are there?

4.1 What is theory?

Theory is a slippery term. It sometimes connotes binary oppositions (theory versus practice, abstract versus concrete, possible but unlikely ('in theory') versus actual or empirical ('in practice'), general versus specific, theoretical versus useful) (cf. Biesta *et al*., 2011; Hammersley, 2012). 'Theory' can mean an opinion or belief ('I have a theory that secondary school males deliberately underperform'), or a tentative explanation ('secondary school males underperform at school in order to look cool in front of their peers'). It can mean a bundle of concepts, for example, a theory of effective schools which includes leadership, achievement orientation, resources, curricular matters, pedagogy and assessment, student motivation, parental support and so on. Such concepts may be related and internally coherent, for example, a theory of 'merit' in a meritocracy which is premised on the combination of IQ and effort. It can mean an explanatory framework, a way of looking at a situation. It can mean law-like statements, for example, 'large organizations have a proclivity to bureaucratization' (though whether there are 'laws' of social behaviour is questionable). It can comprise advocacy or normative principles, for example, 'I have a theory that all children should go to school for a minimum of fifteen years'. It is easy for the educational researcher facing these many interpretations to be lost in a sea of different, indeed contradictory meanings. How, then, can we define 'theory'? In the bullet points that follow, we crystallize several characteristics of 'theory'.

Bacharach (1989) defines theory as 'a statement of relations among concepts within a boundary set of assumptions and constraints' (p. 496), and which approximate to the empirical world (p. 498):

- A theory is a statement, suggestion or proposition that brings together concepts and constructs into a coherent whole, framework or system which has clearly set limits and assumptions.

Huff (2009) notes that theories are explanations of a generalized nature which enable the researcher to compare and analyse empirical data (p. 44). Leong *et al*. (2012) echo this in their comment that theory is 'the story behind the variables … the explanation as to why the variables are related' (p. 122). Hammersley (2012) sets out several meanings of a theory:

> *Theory in relation to practice*. Ideas about how an activity of a particular type *ought to be* carried out, why, what its value is, and so on. On this interpretation, theory is normative in character …
>
> *Theory versus fact*. Sometimes it is said that a particular statement is 'only a theory', implying that it is not well-established knowledge but hypothetical interpretations. Here, theories are factual rather than normative but at the same time speculative in character: their validity is uncertain, or they may even be viewed as idealizations …
>
> *Theory as abstraction as against concrete particulars*.… [T]he distinctive feature of theory here is that it operates at a level of abstraction that is higher than immediate experience or commonsense knowledge …
>
> *Theory as concerned with the macro, as against accounts of the local*. … [T]he term 'theory' being restricted to accounts that have a broad rather than a local focus …
>
> *Theory by contrast with description*. Theories tell us 'what causes what' …

> *Theory as an explanatory language.* Wider than a single explanatory principle … any true theory must be a *set* of principles that tell us about the *whole range of behaviour* of some type of social phenomenon … basic principles of causal systems, these being hidden from ordinary forms of perception and cognition.
>
> *Theory as an approach or 'paradigm'.…* [I]nvolving whole philosophies, in the sense of distinctive sets of ontological, epistemological, and perhaps also praxiological, assumptions.
>
> (Hammersley, 2012, pp. 393–9)

Echoing Bacharach (1989), Alvesson and Sandberg (2013) note that a theory is not 'free-floating' (p. 51) but 'based on and bounded by researchers' assumptions about the subject matter in question' (p. 51). Bacharach argues that theories serve to eliminate or simplify the 'complexity of the real world' (1989, p. 497) but they are not, themselves, data or their categorization, typologies or metaphors. Theories move beyond the 'what' to the 'why', 'how' and 'when' questions (p. 497), i.e. they explain.

Theory is defined by its purposes. How we define theory is made clear by what we want theory to do, the uses to which it is put, for example, to describe, clarify, understand (and more broadly and deeply), make sense of, make intelligible, conceptualize, interpret, explain, predict, generalize, provide answers, empower and emancipate. Definitions of theory are differentiated by its purposes. Biesta *et al*. (2011) remark that we need theory in any sort of educational research, as the essence of research is unavoidably interpretative (p. 230). They give the example of research into 'learning' (p. 233): unless we have a clear concept of what we mean by 'learning' then we cannot usefully research it, and different views of 'learning', for example, as changing behaviour, as processing information, as acquiring knowledge (p. 233), have different areas of focus, methodologies and definitions of what counts as relevant data.

Biesta *et al*. (2011) suggest that one important and enduring quality of theory is to make visible and intelligible those things which might not be so or which might not be immediately able to be observed (p. 227). In an empirical world we need theory to infer causality and causal processes or mechanisms (p. 228), to generate explanations, to give 'plausibility' to explanations (p. 229). As they remark, one task of theory is to enable questions to be answered concerning why people say what they do, do what they do and act as they act (p. 230). Without theory we can only observe correlations, and we need theory to infer or make sense of putative causality. In a hermeneutic world we need theory to understand and interpret experiences, social behaviour, societies, texts and discourses. In the world of critical theory we need theory to interrogate unequal power relations that disfigure lives, to critique inequalities, to emancipate and to transform participants and societies. In other words, theories not only differ by their purposes but by their consequences.

'Theory' has been defined by Kerlinger (1970) as 'a set of interrelated constructs (concepts), definitions, and propositions that presents a systematic view of phenomena by specifying relations among variables, with the purpose of explaining and predicting the phenomena' (p. 9), i.e.:

- A theory specifies the relationship between its elements or component parts, concepts and constructs;
- A theory describes;
- A theory explains;
- A theory predicts.

Theory gathers together all the isolated pieces of empirical data into a coherent conceptual framework of wider applicability. For Bacharach (1989), a theory can be regarded as 'a system of constructs and variables in which the constructs are related to each other by propositions and the variables are related to each other by hypotheses' (p. 498). More than this, however, theory is itself a potential source of further information and discoveries, a source of new hypotheses and hitherto unasked questions; it identifies critical areas for further investigation; it discloses gaps in our knowledge; and it enables a researcher to postulate the existence of previously unknown phenomena.

Theoretical frameworks differ from conceptual frameworks. Conceptual frameworks specify the key concepts being employed in a particular study, how they are used to explore the phenomenon in question, the sequence in which the concepts figure in the research and the direction of relationships of the variables and concepts in the framework. A conceptual framework indicates the relationships of concepts which are concrete and specific to the piece of research in question. By contrast, theoretical frameworks seek to explain and predict, and are at a higher level of abstraction and generality than conceptual frameworks; indeed they appeal to generalizability, which is not the stuff of conceptual frameworks. They are based on the accumulated wisdom of multiple tests and research; they are the general ideas which underpin the conceptual relationships, and are not specific to the study in question. The theory might explain why relationships between

concepts exist, what connects them. Theory is abstract, a generalization that explains relationships between concepts and phenomena.

For example, a researcher looking at learning may have a conceptual framework which suggests the variables involved in understanding how a particular pedagogic strategy improves student's academic achievement. The theoretical framework underpinning this might be, for example, stimulus-response theory, or motivational theory, or self-efficacy theory, or constructivist theory.

A further characteristic of a theory concerns its scope. Whereas an *explanation* might hold for a specific event, situation or issue:

- A theory is a generalized and generalizable statement, i.e. it holds true across contexts beyond those that gave rise to the theory and beyond the specific case in question.

A theory and an explanation are different. Explanations tend to be more specific than theories, or, put another way, theories are more general than explanations, specifying *principles*. Whereas an explanation may focus on a specific case, a theory focuses on *types* of cases or phenomena (Hammersley, 2014, p. 34), drawing on principles; it is independent of the specific case or phenomenon under consideration. Theories hold true beyond the case, population or phenomenon in question (p. 35), and appeal to more general principles and/or causal statements or claims.

A theory lies behind a proposition or hypothesis which is tested, which is falsifiable; it informs, generates, gives rise to a proposition, hypothesis or research question:

- A theory is a general set of principles that are independent of the specific case, situation, phenomenon or observation to be explained.

For example, a theory of effective learning may hold that students learn effectively when non-cognitive factors are included in learning, and this theory may give rise to a hypothesis: 'Students whose intrinsic motivation and self-esteem are high score higher in mathematics than students whose intrinsic motivation and self-esteem are low.' Behind the hypothesis is a framework of coherent, related elements, for example, effective learning is influenced by interrelated non-cognitive elements of personality such as motivation, self-esteem, self-image, disposition and attitudes to learning, interests, sense of responsibility and so on, some of which are included in the hypothesis here.

A theory, standing behind a particular case or phenomenon under investigation, comprises its own set of interrelated constructs or concepts, like Kuhn's (1962) view of a paradigm:

- A theory is a way of looking at and seeing things, conducting research (methodologies, methods and truth tests) and setting research agendas: what has to be researched and how.

For example, I can view the world of schooling through the lens of the hidden curriculum of crowds, praise, power and denial (Jackson, 1968), or I can look at the drive for qualifications through the lens of the credentialist spiral (Oxenham, 1984). These drive what we see, underlining the view that observations are not theory-free (Popper, 1968, 1980). Indeed, observations are inescapably theory-laden in terms of what to look at or for, what not to look at or for, how to look and how to interpret what we see. The theory determines the observation. An artist looking at a rocky mountainous landscape will focus on certain features and see it differently from, say, a mountaineer, a farmer or a geologist. For example, I may investigate the increasingly tight relationship between education and occupational status, to determine whether it is a result of:

- meritocracy and the move away from ascription and towards achievement;
- increased credentialism (qualifications becoming the first filter in job appointments);
- lean-and-mean employment practices (reduced numbers of workers in a company/organization combined with greater demand on those who are employed);
- increased skill-level requirements;
- increased competition for jobs;
- limited employment and career prospects (the supply side);
- increasing demands (the demand side);
- a range of diverse individual motives that are not caught in simple, generalized independent variables.

What I look for, how I look, what evidence I gather and how I interpret the data are determined by the theoretical lenses through which I am looking at the situation.

Theories must be put to the test of rigorous evidence, and we filter out, or do not even consider, what we think are irrelevant factors or data. This is not to say that we cannot proceed in research unless we have deliberately articulated and made explicit our theory; indeed we can conduct educational research without having such articulation. A case can be made for

conducting research without any explicit appeal to a specific theory (e.g. Carr, 2006). As Gorard (2013, p. 31) remarks, 'theory is very much the junior partner in the research process', not least since any case, situation or phenomenon can be explained by any number of theories (Ary *et al.*, 2006). Indeed theories may be unnecessarily and undesirably restricting if we only operate with the delimited boundaries of specific theories when investigating complex, multi-dimensional, multi-perspectival matters.

Hitchcock and Hughes (1995) draw together several strands of the previous discussion in describing a theory thus:

> Theory is seen as being concerned with the development of systematic construction of knowledge of the social world. In doing this theory employs the use of concepts, systems, models, structures, beliefs and ideas, hypotheses (theories) in order to make statements about particular types of actions, events or activities, so as to make analyses of their causes, consequences and process. That is, to explain events in ways which are consistent with a particular philosophical rationale or, for example, a particular sociological or psychological perspective. Theories therefore aim to both propose and analyze sets of relations existing between a number of variables when certain regularities and continuities can be demonstrated via empirical enquiry.
>
> (Hitchcock and Hughes, 1995, pp. 20–1)

Do we need theory in order to conduct educational research? The answer depends on what we mean by 'theory', what kinds of theory we are addressing, what role a theory plays (what we want it to do). The bullet points we have identified so far set out some characteristics of 'theory'.

4.2 Why have theory?

A theory helps us to select, classify and organize ideas, processes and concepts. It helps us to explain, clarify and articulate the heart of the issue. Theory helps us to formulate and find causal relationships; it helps us to understand what, how and why observed phenomena and regularities occur. Theory helps us to predict, for example, outcomes, relationships, and to answer the question 'what will happen if…?' It guides the direction of the research, identifying key fields, methods of working, key concepts; in other words, it serves as a basis for action.

Having a firm theoretical base strengthens research, as it identifies assumptions and enables the researcher to evaluate and critique them. Theory and theoretical frameworks connect the researcher to existing knowledge in the field, are a frame of reference, identify new issues and areas in that field and provide a basis for hypothesis formulation and testing. Theoretical frameworks identify key variables operating in a phenomenon, key concepts and the conceptual basis and framework of the research; they identify and articulate research problems/questions and how to research them. Having a theoretical framework to the research clarifies which facts and evidence will and will not be relevant and important in the research and what are the important research questions that need to be posed to understand and explain an issue. Theory enables the researcher to move to generalization and to identify some of the limits of a generalization.

4.3 What makes a theory interesting?

Educational researchers seeking to ensure that their research is influential, with high impact, should work with 'interesting' theories, i.e. those theories which break new ground or cause us to look at phenomena differently, or discover new features of a phenomenon. Echoing Davis (1971), Alvesson and Sandberg (2013) argue that 'interesting theories' are those that challenge our 'taken-for-granted assumptions in some significant way' (p. 4), that problematize them. In this respect they have higher impact than 'incremental' and 'gap spotting' theories (pp. 4–5), as these latter two tend to work within given agendas rather than to challenge them. Indeed Davis's seminal article indicates twelve ways in which theory can be interesting by challenging the 'taken-for-granted world of their audience' (Davis, 1971, p. 311) (see Morrison and van der Werf (2015) for examples of how Davis's work is illustrated in educational research):

1. *Organization*: What appears to be an unstructured mass of disorganized matter or phenomena is actually the opposite, and vice versa (p. 311).
2. *Composition*: What appear to be assorted matters or phenomena are actually a single matter or phenomenon, and vice versa (p. 315).
3. *Abstraction*: What appears to be an individual matter or phenomenon is actually a holistic matter or phenomenon, and vice versa (p. 316).
4. *Generalization*: What appears to be a local matter or phenomenon is actually a general matter or phenomenon, and vice versa (p. 317).
5. *Stabilization*: What appears to be a stable and immutable matter or phenomenon is actually the opposite, and vice versa (p. 318).

6 *Function*: What appears to be something that works poorly in achieving an aim is actually the opposite, and vice versa (pp. 319–20).
7 *Evaluation*: What appears to be a bad matter or phenomenon is actually the opposite, and vice versa (p. 321).
8 *Co-relation*: What appear to be unrelated items or phenomena are actually the opposite, i.e. are interdependent, and vice versa (p. 322).
9 *Coexistence*: What appear to be matters or phenomena that can coexist actually cannot, and vice versa (pp. 23–4).
10 *Co-variation*: What appear to be positive co-variations between matters or phenomena actually are negative co-variations, and vice versa (p. 324).
11 *Opposition*: What appear to be identical or very similar matters or phenomena are actually the opposite, and vice versa (p. 325).
12 *Causation*: What appears to be an independent causal variable is actually a dependent variable, and vice versa (p. 326).

Alvesson's and Sandberg's (2013) methodology for generating 'interesting' research requires researchers to expose and evaluate assumptions (e.g. in the literature, in 'theories'), and, from there, to develop and evaluate an alternative ground of assumptions (p. 56). Some 'interesting' theories act as a bridge between two or more different theories (Bacharach, 1989, p. 511) or identify redundant or incorrect earlier theories, i.e. help us to re-evaluate existing theories. As Kaplan (1964) remarks: a 'new theory requires its own terms and generates its own laws: the concepts are not merely reorganized, but reconstituted, the old laws are not just connected but given a new meaning' (p. 297).

How we conduct research is informed by the *types* of theory in which we are working or which underpin the research, and these are introduced below.

4.4 Types of theory

There are several different types of theory, and each type of theory defines its own kinds of 'proof': for example, *empirical theory*, '*grand*' *theory*, *normative theory* and *grounded theory* (discussed below), and '*critical*' *theory* (introduced below as an instance of normative theory and which has the entire Chapter 3 devoted to it). The status of theory varies quite considerably according to the discipline or area of knowledge in question. Some theories, as in the natural sciences, are characterized by a high degree of elegance and sophistication; others, perhaps like educational theory, are only at the early stages of formulation and are thus characterized by unevenness.

Empirical theories

Many definitions of 'theory', which add to the features of 'theory' set out in the bullet points earlier, include a requirement that locates it in an empirical context, relating to observational evidence (widely defined), of which scientific theories are prime examples:

- A theory is based on, and guides, empirical research, typically in hypothesis testing or testing by making systematic, objective predictions and observations;
- A theory is testable and, therefore, falsifiable and provisional, such as a scientific theory;
- A theory is conjectural, suggesting relations (e.g. correlational or causal) between two or more items (however defined, for example, variables, factors, observations).

Kettley (2012) suggests that an empirical theory is a 'coherent description and explanation of observed phenomena which provides a testable, verifiable or falsifiable, representation of social relationships' which 'enables the researcher to speculate about future social activity and, perhaps, to predict behaviour drawing on the inferences of the explanation' (p. 9). Gorard (2013), too, suggests that a theory 'is a tentative explanation' (p. 31). He holds that a reasonable theory is one that offers a reasonable explanation based on the research evidence, that this is surpassed by a theory which not only explains the observations but holds true when tested further, and, in turn, this is surpassed by a theory which not only explains and survives further testing, but which has predictive value for something which is completely unexpected (p. 31).

Here a theory has a tentative, impermanent and conjectural quality; it can be tested, upheld, disproved or modified, and its strength resides in its surviving such 'severe tests', i.e. those tests which are deliberately designed to try to falsify the theory (Popper, 1968, 1980). Popper's (1968) view of a scientific theory takes the form of a universal law applying to a particular type of phenomenon. Such a law should demonstrate precision and universality, 'it should set the criteria for its own falsification' (p. 92) and possess explanatory and predictive power. Indeed Popper (1968) comments that the 'best theory' is that which is testable, survives being tested, has greater explanatory power than competing theories and has the greatest content and simplicity (p. 419).

Popper (1968), Lakatos (1970), Mouly (1978), Laudan (1990) and Rasmussen (1990) identify the following

characteristics of a sound empirical theory (cf. Morrison, 1995a). It should:

- be operationalizable precisely;
- be testable, and against evidence which is different from that which gave rise to the theory, i.e. moving beyond simply corroboration and induction and towards 'testing', identifying the type of evidence which is required to confirm or refute the theory;
- permit deductions to be made;
- have explanatory and causal power;
- be compatible with both observation and previously validated theories. It must be grounded in empirical data that have been verified and must rest on sound postulates and hypotheses. The better the theory, the more adequately it can explain the phenomena under consideration, and the more facts it can incorporate into a meaningful structure of ever-greater generalizability. There should be internal consistency between these facts;
- clarify the precise terms in which it seeks to explain, predict and generalize about empirical phenomena;
- be tentative, conjectural, provisional and falsifiable, stating the grounds, criteria and circumstances for its own empirical verification, proof, falsification or rejection;
- demonstrate precision and universality, identifying the nature and operation of a 'severe test' (Popper, 1968), permit deductions that can be tested empirically, i.e. it must provide the means for its confirmation or rejection. One can test the validity of a theory through the validity of the propositions (hypotheses) that can be derived from it. If repeated attempts to disconfirm its various hypotheses fail, then greater confidence can be placed in its validity. This can go on indefinitely;
- clarify its methodologies (e.g. hypothetico-deductive, inductive);
- be faithful to the subject matter and evidence/data from which it has been derived;
- be useful: describe and explain all the relevant data and observations;
- be clear: clarify the conceptual framework and the paradigm in which it works;
- demonstrate internal coherence of its component elements, consistency and internal logic;
- have great explanatory, predictive, retrodictive and generalizable potential;
- be replicable;
- have logical and empirical adequacy and internal coherence;
- be able to respond to observed anomalies;
- be parsimonious, excluding any unnecessary ideas and explanations and stated in simple terms; that theory is best which explains the most in the simplest way. A theory must explain the data adequately and yet must not be so comprehensive as to be unwieldy. On the other hand, it must not overlook variables simply because they are difficult to explain;
- be corrigible in light of further evidence;
- be a spur to empirical research, spawning research and new ideas (fertility);
- lead to new ideas that would otherwise not have emerged.

Empirical theories, by their very nature, are provisional. A theory can never be complete in the sense that it encompasses all that can be known or understood or certain about the given phenomenon. As Mouly (1978) argues, one (scientific) theory is replaced by a superior, more sophisticated theory, as new knowledge is acquired (echoing Kuhn's (1962) discussion of paradigms and paradigm shifts). An empirical theory gathers together all the isolated bits of empirical data into a coherent conceptual framework of wider applicability. More than this, however, empirical theory is itself a potential source of further information and discoveries. In this way it is a source of new hypotheses and hitherto unasked questions; it identifies critical areas for further investigation; it discloses gaps in our knowledge; and enables a researcher to postulate the existence of previously unknown phenomena.

Grand theory

Not all theories are testable. Some theories are artefacts comprising abstract concepts, for example, theories of modernism and postmodernism, Freudian theory, Talcott Parsons's theory of social formations and systems, Habermas's theory of communicative action, Bourdieu's theory of *habitus*. Such theories are very different from testable theories or hypotheses, and often include large-scale conceptual frameworks which are used to comment on or understand phenomena or explain them in broad terms ('grand theory'), and are not bounded by space and time (Bacharach, 1989, p. 500).

Such 'grand theory' (from Mills's *The Sociological Imagination*, 1959) is typically abstract and removed from a specific situation, standing back to set out a view of the world through a conceptual framework. It can be formal, conceptual, speculative, overarching and non-empirical, and it defines areas of study, clarifies and refines their conceptual frameworks and enlarges the way we consider the social and educational world. Grand theory sets out some fundamental ontological and epistemological frameworks and concepts which

define an area of study or domain of enquiry (cf. Layder, 1994, pp. 28–30). Grand theory is untestable and unable to be 'proved' or 'disproved', being a way of considering the world, and, in that respect, is more like an orientation, a rationalization, a belief or an article of faith (cf. Gorard, 2013, p. 32) than a testable, scientific theory. It is a discourse.

Grand theory is a metanarrative, defining an area of study, being speculative, clarifying conceptual structures and frameworks and enlarging the way we consider phenomena (Layder, 1994). It defines a field of enquiry (Hughes, 1976) and uses empirical material by way of illustration rather than 'proof' (p. 44). This is the stuff of some sociological theories, for example Marxism, rational choice theory, structuralism and functionalism.

Whilst sociologists may be excited by the totalizing and all-encompassing nature of such grand theories, they have been subject to considerable critique. For example, Merton (1957), Coser and Rosenberg (1969), Doll (1993) and Layder (1994) contend that whilst they might possess the attraction of large philosophical systems of considerable – Byzantine – architectonic splendour and logical consistency, nevertheless they are scientifically sterile, irrelevant and out of touch with a world that is characterized by openness, fluidity, change, heterogeneity and fragmentation.

However, Murphy (2013) makes a strong case for the applicability of such social theories to everyday practices and lives, as they can explain: social change and development; how and why people behave as they do; the operations of power, culture and social structures; issues of gender, race, class, ability and identity; modernity and postmodernity; institutions and their operations, and so on. He argues powerfully for the role of social theory in educational research, indeed his website for Social Theory Applied (http://socialtheoryapplied.com) provides many links between theory, practice and research. For example, in the field of education he argues that social theory can inform topics such as inequality and inclusion, educational selves and subjectivities, curricular and pedagogic practice, and governance and management (pp. 8–9), and he draws on the social theories of Habermas, Foucault, Bourdieu and Derrida.

Grand theories have been criticized for their aridity and inability to stand empirical scrutiny or testing (Merton, 1957; Mills, 1959; Layder, 1994). This charge, however, might appear unfair, attempting to judge a theory by criteria which it does not strive to meet. There remains the problem that too easily grand theory can become empty rationalization; for example, Hughes (1976) comments that 'this form of theorizing is just so much over-elaboration of concepts almost to the wilful exclusion of any empirical import' (p. 45) and that it involves elaborating distinctions arbitrarily which do little or nothing to increase understanding or make greater sense of experience (Mills, 1959, p. 33). This echoes Merton's (1957) view that, though they may have the architectonic splendour of 'large philosophical systems', they are also marked by 'scientific sterility' (p. 10) and, whilst being admirable for their logical consistency, are largely of no relevance to the everyday world (Coser and Rosenberg, 1969, p. 14). To add to this is the familiar critique of grand theory as being totalizing metanarratives in a world in which no single metanarrative is operable.

For the educational researcher, grand theories can inform an understanding of the world and articulate a way of looking at phenomena or explain the context of a study, and, in this respect, might prompt the development of research questions (cf. White, 2009, p. 26). Whereas empirical theory looks to 'proof' as a criterion of its validity, grand theory looks to logical coherence, explanatory potential and articulation of key concepts in understanding a phenomenon.

Middle-range theory

Between 'grand' theory and small-scale theories or minor working hypotheses lie what Merton (1967) termed 'middle-range theories': subsets of overarching theories which focus on specific topics and seek to explain them. These typify much educational research. Middle-range theory focuses on a particular phenomenon or case in context, and seeks to explain it in terms of underlying mechanisms, factors or principles that give rise to the phenomenon or case in point. It uses a limited set of assumptions to derive hypotheses/questions logically which can be tested empirically (Merton, 1957). It starts with a specific empirical phenomenon (in contrast to a broad abstract area such as 'capitalism' or 'social structure', which is the stuff of 'grand theory') and abstracts and creates from the phenomenon general statements which can be verified by data.

Merton (1967), in defining middle-range theories, notes that this intermediate position uses abstractions and concepts 'close enough to observed data to be incorporated in propositions that permit empirical testing' (p. 39) in studying delimited aspects of social life. Here 'theory' comprises propositions which are logically interconnected and from which 'empirical uniformities' can be derived to explain all the observed uniformities of social behaviour, social organization and social change (p. 39). He is against theories that are so abstract that they cannot be tested, but he also argues

for middle-range theories which can apply abstract concepts to different spheres of social structure and social behaviour, i.e. which bridge micro- and macro-social problems, looking at here-and-now matters.

Pawson (2008) provides an example of middle-range theory in education. In improving higher education, recourse may be made to a variant of 'naming and shaming', based on the theory that publishing comparative data on higher education institutions (e.g. rankings) will stimulate competition between them, and thereby drive up standards (p. 18).

Middle-range theory draws on empirical evidence, and embraces, but does not confine itself to, hypothesis testing, prediction, isolation and control of variables. It can be explanatory and interpretive, seeking to understand a situation in its context, a hermeneutic, practical exercise, following the *verstehen* approaches of Weber, looking at meaning in social contexts. Much educational research appeals to middle-range theory, focusing on a specific case or phenomenon and seeking to explain it using concepts and hypotheses which, whilst being somewhat abstracted from the specific case in question, are not part of the parlance of grand, totalizing theoretical edifices.

Normative theory

Normative theories explain how people, groups, institutions etc. *ought* to operate within a specific system of social values (norms). They are prescriptive, for example: 'all students should be taught ideology critique'; 'schools should promote democracy'.

A clear example of a normative theory is critical theory, which is such a large field that we devote an entire chapter to it (Chapter 3). Critical theory seeks to uncover the interests at work in particular situations, to interrogate the legitimacy of those interests and to identify the extent to which they are legitimate in serving equality and democracy. It has a deliberate intention of being *transformative* and *emancipatory*, promoting democracy and individual freedoms. It is practical and political, to bring about a more just, egalitarian society in which individual and collective freedoms operate, and it seeks to eradicate the exercise and effects of illegitimate power.

Critical theory, operating through ideology critique, identifies unequal power relations in society, interrogates their legitimacy, identifies what has brought an individual or social group to relative powerlessness or, indeed, to power, and questions the legitimacy of repression, voice, ideology, power, participation, representation, inclusion and interests. It argues that much behaviour (including research behaviour) is the outcome of particular illegitimate, dominatory and repressive factors, illegitimate in the sense that they do not operate in the general interest – one person's or group's freedom and power is bought at the price of another's freedom and power.

For the educational researcher, a normative theory, of which critical theory is an example, should be clear in its methodology, set criteria for its validation (both empirical and non-empirical) and denote the type(s) of evidence that could substantiate the theory. As with other views of 'theory', normative theory must have substantive concepts which are internally coherent and logically tenable. It might be a middle-range theory, 'grand theory' or empirical theory. It should demonstrate appropriate precision and universality, possess explanatory power and predictive validity, and, as for other theories, have greater validity claims and warrants than rival theories. Criteria for judging its worth include its ability to achieve the norms and values explicit in the theory, for example, for critical theory this means its potential for, and achievement of, practical empowerment, freedom, equality, social justice, democracy and emancipation.

A contentious issue raised by normative theory is whether it is the task of educational research to have an explicit political or ideological agenda, to engage in political activism and/or policy making, or whether educational research should simply stick to providing factual knowledge that is used by others for political, normative agendas and policy making (Hammersley, 2014). Should educational researchers be concerned only with the neutral, disinterested pursuit and provision of knowledge (recognizing that this is itself a value position) or seek to press home a particular ideology or set of values? Should educational researchers directly answer questions of values, of desirability, of right and wrong, or should they just stick to facts, not values? We discuss this in Chapter 3.

Grounded theory

Grounded theory (addressed fully in Chapter 37) is not predetermined, but, rather, emerges from, and is consequent to, data (Glaser and Strauss, 1967), i.e. it is grounded in data and rises up from the ground of data: a 'bottom-up' process. It seeks to *generate* rather than simply to test an existing theory. Grounded theory commences with data on a topic or phenomenon of concern, and then, using tools such as theoretical sampling, coding, constant comparison, identification of the core variable, and saturation, the theory – the explanation and explanatory framework – emerges from the analysis and study of, and reflection on, the phenomena and data under scrutiny (cf. Strauss and Corbin, 1990, p. 23). Grounded theory identifies key features and

relationships emerging from data and categories and then proposes a plausible explanation of the phenomenon under study by drawing on the data generated, ensuring that the theory explains all the data without exception.

A concern of researchers working with grounded theory is how far one can generalize from it: is the grounded theory limited to the specific study; can it be applied to other similar situations; does it have wider generalizability like a 'grand' theory? We explore this in Chapter 37, noting that grounded theory researchers often refer to 'transferability' of the findings from one situation to another, which is based on the judgement of the researcher or reader, and that it is problematic to rely too heavily on reader judgement in determining the status of the grounded theory. Grounded theory has some affinity to 'middle-range' theories.

Empirical, grand, middle-range, normative and grounded theories have practical value. They underpin research design, data analysis and, indeed, the generation of new theories. For example, they suggest the relationships between variables and concepts; they can be used to set up the research, what it seeks to do and what key concepts are included in the research; they can predict and explain findings, suggesting at a high level of abstraction *why* such and such occurs.

Social theories and psychological theories can be usefully brought into educational research. We give a fully worked example in Chapter 6, and we refer readers to this. By way of summary of that example here, Goldthorpe (2007) plans, conducts and reports research which explains the causes of 'persistent differentials in educational attainment' despite increased educational expansion, provision and uptake across the class structure (p. 21). In his research, theory plays a part at each of seven stages:

Stage 1: Establish exactly what has to be explained, examining regularities and patterns of relevant phenomena.
Stage 2: Set out possible theoretical foundations for the investigation, which utilize high-level sociological and economic theory (Marxist theory, liberal theory, cultural theory and rational choice theory).
Stage 3: Examine, evaluate and eliminate rival theoretical foundations, selecting the most fitting and justifying the selection (in his research it is rational choice theory).
Stage 4: Hypothesize a causal explanation on the basis of the best theoretical foundation. This operates at a sophisticated theoretical level, arguing that different classes view the costs, risks and benefits differently (p. 34).
Stage 5: Set out the assumptions underlying the causal explanation (which concerns income differentials, class differentials, risk aversion, anticipated costs and benefits).
Stage 6: Test the causal hypotheses empirically, in which data are collected on differences in aspirations and decisions by social class which are caused by: (a) perceptions of costs; (b) relative risk aversion; and (c) perceptions of relative benefit.
Stage 7: Draw conclusions based on the test. Here, there are class differences in terms of relative ambition, risk aversion, perceived costs and benefits, amounts of effort required, assurances of success (and the significance of this), fear of downward social mobility, income, occupational choices, and the need for qualifications. These, in turn, based on empirical data, support his explanation of the factors of relative risk aversion and fear of downward social mobility exerting causal power on educational decision making which, in turn, lead to class differentials in educational attainment being maintained (p. 99).

In this example, grand theory (sociological theory of rational choice), empirical theory (e.g. hypothesis generation and testing) and middle-range theory (the explanation of a particular phenomenon) all have their place at different stages of the research. We advise readers to look at the example in Chapter 6 for a fuller account of this.

4.5 Where does theory come from?

Where does a theory come from? How does one establish a theory? Clearly there are many starting points. One starting point may be through observation and analysis, for example, of observed regularities or relationships, of an association of events, of data; another is through reflection and creativity. Another may be through asking a 'what if' question, for example, 'if assessment were to become more authentic, would it increase student motivation?', or 'does repeating a year at school improve student performance?' Another starting point might be less concerned with regularities than a single instance: 'why did such-and-such happen?' or 'why is such-and-such happening?' Another starting point may be from literature which gives rise to a theory, or previous research.

Echoing C. Wright Mills's (1959) view of *The Sociological Imagination*, theory generation is a human act. It is the creative imagination in the minds of humans which links concepts and sets the grounds for logical coherence and theory validation. As Bacharach (1989) argues, theories derive from people's 'creative imagination and

ideological orientation or life experience' (p. 498). Theory comes from humans creating and connecting ideas and concepts into an explanatory framework within articulated boundaries, and testing them.

Similarly, it is humans who decide whether a theory holds water; who proffer explanations, predictions and generalizations; who select what is relevant to include in a theory; and who determine how to test, validate and falsify the theory. Whilst observations and data may provide the fuel for theory generation and testing, they are not the theory itself.

Theory typically precedes research questions; research does not start with research questions. We assemble observations, ideas, concepts, reflections, consider what they mean, and then formulate our theories, our frameworks of related concepts and propositions. Then we construct our tests of the theories, which may utilise a hypothetico-deductive empirical method (e.g. correlational, causal analysis, difference testing, regressions, or other kinds of analysis and intervention), a hermeneutic method, an emancipatory, transformative method through ideology critique, a grounded theory approach, or others.

Depending on the type of research and research question, we often commence with a theory and then test it, moving from theory to hypothesis generation to hypothesis testing and observation to prediction to conclusion to generalization. Alternatively, as in a grounded theory approach, we may conduct post hoc theory generation, i.e. starting with data and, through the tools of grounded theory, end up with an emergent theory which subsequently we may wish to test in other contexts and conditions.

4.6 Questions about theory for researchers

In considering the role of theory in educational research, researchers can address the following questions:

- What definition of theory are you using?
- What is your theory (state it clearly)? Is it a hypothesis, a set of related concepts, a value system, a political/ideological agenda, an explanatory framework, a possible explanation, an opinion, an approach etc.?
- What is the theory/theoretical framework in which you are working? What are its key components, constructs, concepts and elements, and how do they relate to each other logically and coherently?
- What is your theory seeking to describe ('what'), explain ('how', 'why', 'when'), predict ('what if'), generalize, i.e. what is it a theory of, and why is this relevant for your research?
- What makes your theory interesting?
- How important is theory in your research? (Why) do you need to make it explicit?
- What theories are you using, and why these: how relevant are they to your study?
- What is the purpose of your research with regard to theory, for example, to test, apply, explain, understand, generate, critique, validate, extend, refine, refute a theory?
- What is the relationship between your theory, your research and your research question(s)?
- What type of theory are you using (e.g. empirical, grand, normative, middle-range, critical, grounded)?
- What methodologies are you using to work with your theory (e.g. empirical testing, hermeneutic interpretation and explanation, ideology critique)?
- What criteria are you using to validate your theory (e.g. compatibility with empirical data, logical coherence and adequacy, explanatory potential, achievement of transformative and emancipatory potential etc.)? How will you validate your theory?
- How does your theory give rise to testable propositions/hypotheses, or inform a hermeneutic exercise, or bring about its espoused values or normative intentions?
- What are the boundaries of, and assumptions in, your theory?

4.7 Conclusion

Researchers frequently pose the question 'do we need theory?' This is an inappropriate question, for, one the one hand, like it or not, we cannot escape theory: it is there, it underpins what we do, whether or not we are conscious of it. One of the contributions of the post-positivists is in drawing attention to the point that no observation is theory-free. A more useful question is what we need theory for, as, by itself, it may underdetermine or unnecessarily constrict the full gamut of the research enterprise.

Theories help us to think. They articulate and organize ways of approaching a problem or phenomenon. They assemble and clarify key concepts and their relationships, principles and abstractions, explanations and propositions. They can stimulate research questions and hypotheses. Theories connect concepts into a logical and coherent whole or framework.

Theories help us to learn: they can render ideas testable, define ways of working, tell us which ideas, statements, conclusions, lines of reasoning stand fast when tested rigorously and which appear to be valid, reliable,

credible, legitimate, sound, reasonable and useful. Some theories (e.g. descriptive, analytical, explanatory) help us to understand a phenomenon; others define an approach, conceptual frame and reference system (e.g. grand theory); yet others (e.g. normative, critical) seek to change the world or to promote an agenda. Theories are tools for thinking, describing, understanding, predicting, explaining, proving, organizing, connecting ideas and concepts, generalizing, generating research enterprises and suggesting research questions and answers. It is for each researcher to decide which meaning(s) of 'theory' is/are being used in a specific research project.

Like it or not, use them or not, one or more theories lie behind an educational research study. Whether these drive the research, are incidental or unimportant to it are matters for each researcher. Perhaps middle-range theory is both a useful compromise and, more positively, a useful way forward.

The companion website to the book provides PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: www.routledge.com/cw/cohen.

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Evaluation and research

CHAPTER 5

This brief chapter indicates some key similarities and differences between research and evaluation. The chapter sets out:

- similarities between research and evaluation
- differences between research and evaluation
- connections between evaluation, research, politics and policy making

5.1 Similarities and differences between research and evaluation

There are many similarities between research and evaluation. Both share commonalities in terms of methodologies, ethical issues, sampling, reliability and validity, instrumentation and data analysis, i.e. their operational practices (Arthur and Cox, 2014, p. 139; Garcia *et al*., 2014). However, there are also differences; the problem of trying to identify differences between evaluation and research is compounded because not only do they share several of the same methodological characteristics, but one branch of research is called *evaluative research* or *applied research*. This is often kept separate from 'blue-skies' research in that the latter is open-ended, exploratory, contributes something original to the substantive field and extends the frontiers of knowledge and theory, whereas in the former the theory is *given* rather than *interrogated* or *tested*. Plewis and Mason (2005, p. 192) suggest that evaluation research is, at heart, applied research that uses the tools of research in the social sciences to provide answers to the effectiveness and effects of programmes. One can detect many similarities between the two in that they both use methodologies and methods of social science research generally (Norris, 1990), covering, for example:

- the need to clarify the purposes of the investigation;
- the need to operationalize purposes and areas of investigation;
- the need to address principles of research design that include:
  - **a** formulating *operational questions*,
  - **b** deciding appropriate *methodologies*,
  - **c** deciding which *instruments* to use for data collection,
  - **d** deciding on the *sample* for the investigation,
  - **e** addressing *reliability* and *validity* in the investigation and instrumentation,
  - **f** addressing *ethical* issues in conducting the investigation,
  - **g** deciding on *data-analysis* techniques,
  - **h** deciding on *reporting* and *interpreting* results.

The features outlined above embrace many elements of the scientific method (see Chapter 1).

Researchers and evaluators pose questions and hypotheses, select samples, manipulate and measure variables, compute statistics and data, and state conclusions (cf. Garcia *et al*., 2014). Nevertheless there are important differences between evaluation and research that are not always obvious simply by looking at publications. Publications do not always make clear the background events that gave rise to the investigation, nor do they always make clear the uses of the material that they report, nor do they always make clear what the dissemination rights are (Sanday, 1993) and who holds them. Several commentators set out some of the differences between evaluation and research. For example, Smith and Glass (1987) offer eight main differences:

1. *The intents and purposes of the investigation*: the researcher wants to advance the frontiers of knowledge of phenomena, to contribute to theory and to be able to make generalizations; the evaluator is less interested in contributing to theory or the general body of knowledge. Evaluation is more parochial than universal (pp. 33–4).
2. *The scope of the investigation*: evaluation studies tend to be more comprehensive than research in the number and variety of aspects of a programme that are being studied (p. 34).
3. *Values in the investigation*: much research aspires to value-neutrality. Evaluations must represent multiple sets of values and include data on these values.

4 *The origins of the study*: research has its origins and motivation in the researcher's curiosity and desire to know (p. 34). The researcher is autonomous and answerable to colleagues and scientists (i.e. the research community), whereas the evaluator is answerable to clients and stakeholders. The researcher is motivated by a search for knowledge; the evaluator is motivated by the need to solve problems, allocate resources and make decisions. Research studies are public; evaluations are for a restricted audience.
5 *The uses of the study*: the research is used to further knowledge; evaluations are used to inform decisions.
6 *The timeliness of the study*: evaluations must be timely; research need not be. Evaluators' timescales are given; researchers' timescales need not be given.
7 *Criteria for judging the study*: evaluations are judged by the criteria of utility and credibility; research is judged methodologically and by the contribution that it makes to the field (i.e. internal and external validity).
8 *The agendas of the study*: an evaluator's agenda is given; a researcher's agenda is her own.

Norris (1990) reports work by Glass and Worthen in which they identified eleven main differences between evaluation and research:

1 *The motivation of the enquirer*: Research is pursued largely to satisfy curiosity; evaluation is undertaken to contribute to the solution of a problem.
2 *The objectives of the search*: Research and evaluation seek different ends. Research seeks conclusions; evaluation leads to decisions.
3 *Laws versus description*: Research is the quest for laws (nomothetic); evaluation merely seeks to describe a particular thing (idiographic).
4 *The role of explanation*: Proper and useful evaluation can be conducted without producing an explanation of why the product or project is good or bad or of how it operates to produce its effects.
5 *The autonomy of the enquiry*: Evaluation is undertaken at the behest of a client; researchers set their own problems.
6 *Properties of the phenomena that are assessed*: Evaluation seeks to assess social utility directly; research may yield evidence of social utility but often only indirectly.
7 *Universality of the phenomena studied*: Researchers work with constructs having a currency and scope of application that make the objects of evaluation seem parochial by comparison.
8 *Salience of the value question*: In evaluation, value questions are central and usually determine what information is sought.
9 *Investigative techniques*: While there may be legitimate differences between research and evaluation methods, there are far more similarities than differences with regard to techniques and procedures for judging validity.
10 *Criteria for assessing the activity*: The two most important criteria for judging the adequacy of research are internal and external validity; for evaluation they are utility and credibility.
11 *Disciplinary base*: The researcher can afford to pursue inquiry within one discipline; the evaluator cannot.

However, we include below a more comprehensive set of distinguishing features, but it must be emphasized that they are not at all as rigidly separate as the bullet points below might suggest:

- *Origins*: Research questions originate from scholars working in a field; evaluation questions issue from stakeholders.
- *Audiences*: Evaluations are often commissioned and they become the property of the sponsors and are not for the public domain; research is disseminated widely and publicly.
- *Purposes*: Research contributes to knowledge in the field, regardless of its practical application, and provides empirical information, i.e. 'what is'; evaluation is designed to use that information and those facts to judge the worth, merit, value, efficacy, impact and effectiveness of something (Scriven, 2004; Arthur and Cox, 2014), i.e. 'what is valuable' (Mathison, 2007, p. 189). Research is conducted to gain, expand and extend knowledge; evaluation is conducted to assess performance and to provide feedback (Levin-Rozalis, 2003). Research is to generate theory; evaluation is to inform policy making (Patton, 2002). Research is to discover; evaluation is to uncover (Arthur and Cox, 2014). Research seeks to predict what will happen; evaluation concerns what has happened or what is happening (ibid.).
- *Stance*: The evaluator is reactive (e.g. to a programme); the researcher is active and proactive (Levin-Rozalis, 2003, p. 15).
- *Status*: Evaluation is a means to an end; research is an end in itself (Levin-Rozalis, 2003).
- *Focus*: Evaluation is concerned with how well something works; research is concerned with how something works (Mathison, 2007).

- *Outcome focus*: Evaluation is concerned with the achievement of intended outcomes; research may not prescribe or know its intended outcomes in advance (science concerns the unknown).
- *Participants*: Evaluation focuses almost exclusively on stakeholders; research has no such focus (Mathison, 2007).
- *Scope*: Evaluations are concerned with the particular, for example, a focus only on specific programmes. They seek to ensure internal validity and often have a more limited scope than research. Research often seeks to generalize (external validity) and, indeed, may not include evaluation (Priest, 2001).
- *Setting of the agenda*: The evaluator works within a given brief; the researcher has greater control over what will be researched (though often constrained by funding providers). Evaluators work within a set of 'givens', for example, programme, field, participants, terms of reference and agenda, variables; researchers create and construct the field.
- *Relevance*: Relevance to the programme or what is being evaluated is a prime feature of evaluations; relevance for researchers has wider boundaries (e.g. in order to generalize to a wider community). Research may be prompted by interest rather than relevance. For the evaluator, relevance has to take account of timeliness and particularity (Levin-Rozalis, 2003, pp. 20–1).
- *Time frames*: Evaluation begins at the start of the project and finishes at its end; research is ongoing and less time-bound (Levin-Rozalis, 2003) (though this may not be the case with funded research).
- *Uses of results*: Evaluation is designed to improve; research is designed to demonstrate or prove (Stufflebeam, 2001). Evaluation 'provides the basis for decision making; research provides the basis for drawing conclusions' (Mathison, 2007, p. 189). Evaluations might be used to increase or withhold resources or to change practice; research provides information on which others might or might not act, i.e. it does not prescribe.
- *Decision making*: Evaluation is used for micro decision making; research is used for macro decision making (Mathison, 2007, p. 191).
- *Data sources and types*: Evaluation has a wide field of coverage (e.g. costs, benefits, feasibility, justifiability, needs, value for money), so evaluators employ a wider and more eclectic range of evidence from an array of disciplines and sources than researchers.
- *Ownership of data*: The evaluator often cedes ownership to the sponsor, upon completion; the researcher holds onto the intellectual property.
- *Politics of the situation*: The evaluator may be unable to stand outside the politics of the purposes and uses of, or participants in, an evaluation; the researcher provides information for others to use.
- *Use of theory*: Researchers base their studies in social science theory; this is not a necessary component of evaluation (Scriven, 1991). Research is theory-dependent; evaluation is 'field-dependent', i.e. not theory-driven but derived from the participants, the project and stakeholders. Researchers create the research findings; evaluators may (or may not) use research findings (Levin-Rozalis, 2003, pp. 10–11).
- *Reporting*: Evaluators report to stakeholders/commissioners of research; researchers may include these and may also report more widely, for example, in publications (Beney, 2011).
- *Standards for judging quality*: Judgements of research quality are made by peers; judgements of evaluation are made by stakeholders (Patton, 2014). For researchers, standards for judging quality include validity, reliability, accuracy, causality, generalizability, rigour; for evaluators, to these are added utility, feasibility, involvement of stakeholders, side effects, efficacy, fitness for purpose (though, increasingly, utility value and impact are seen as elements for judging research) (Patton, 1998).

Mathison (2007) comments that the two major dimensions for distinguishing between research and evaluation are on the particularization/generalization continuum and the decision-oriented/conclusion-oriented continuum.

The statements above are set out in an either/or manner for conceptual clarity. However, the reality of the situation is nowhere near as clear as this; research and evaluation are not mutually exclusive binary oppositions, nor, in reality, are there differences between them. Their boundaries are permeable, similarities are often greater than differences and there is often overlap; indeed, evaluative research and applied research often bring the two together (Levin-Rozalis, 2003, p. 3); indeed Arthur and Cox (2014) note how easily the two have elided in research and assessment exercises. For each of the above there are many exceptions. For example, both evaluation and research might be concerned with generalization, or, indeed with the particular; evaluation may not be for decision making whereas research may be precisely for this purpose (Mathison, 2007). Both research and evaluation are concerned to produce information and to promote explanation and understanding, both of which can contribute to decision making and policy formation, i.e. both are intimately concerned with politics and both involve political processes, with differences between them being more

matters of degree, and both can operate at different levels: individual, local and institutional to national and international (Arthur and Cox, 2014). Elliott (1991) notes that evaluation is an essential ingredient of action research. Patton (1998) argues that differences between research and evaluation are often arbitrary, and cases can be made for saying that they are the same or, indeed, are different.

Mathison (2007) reports that some people put research as a subset of evaluation whilst others put evaluation as a subset of research (p. 189). MacDonald (1987) argues that '[t]he danger therefore of conceptualizing evaluation as a branch of research is that evaluators become trapped in the restrictive tentacles of research respectability.… How much more productive it would be to define research as a branch of evaluation' (p. 43).

A clue to some of the differences between evaluation and research can be seen in the definition of evaluation. Most definitions of evaluation include reference to several key features: (1) answering specific, given questions; (2) gathering information; (3) making judgements; (4) taking decisions; (5) addressing the politics of a situation (Morrison, 1993, p. 2). Morrison provides one definition of evaluation as: 'the provision of information about specified issues upon which judgements are based and from which decisions for action are taken' (p. 2).

5.2 Evaluation research and policy making

In an era in which educational innovations come thick and fast, often on the crest of one wave after another of government policies and interventions, evaluation has to take to task any notion that quick fixes from interventions 'work' straightforwardly, or at all. This runs counter to claims that slogans of evidence-based practice and 'what works' are unproblematic. Rather, identifying 'what works' contains myriad complexities and challenges. Understanding 'what works' in education recognizes that an intervention often unfolds in unexpected ways because of the complex interplay of participants and contingencies. Often as not, this may frustrate the clean, antiseptic world of policy makers. Innovations rarely 'work' entirely in the ways in which they are intended. A range of factors operate on the situation, and impact, effectiveness and efficacy are viewed differently by different participants. Hence key questions for evaluation concern, for example, which part(s) of an intervention or programme are working well/less well, why, for whom, under what contingencies and conditions, in what circumstances and over what time period (cf. Pawson, 2013, p. 167).

Take, for example, the 'healthy attachments' programme in a range of UK schools, which Pawson (2013) reports. This project was designed to improve students' 'well-being, sense of security and positive regard' (p. 73), which, in turn, sought to reduce risks to health associated with tobacco, alcohol and illegal drugs (p. 73). Pawson, carefully dissecting out the claims made for the intervention, shows that it is typical of 'black-box analysis' (p. 74) and that, actually, the positive results overlooked important elements and were far less clear than the claims made for the project's success. Why? Because a range of events and factors occurred, and people's perspectives and situations were neglected in the evaluation which, if taken more seriously, would have led to much more cautious claims. As he says, in response to the question 'is this the correct interpretation?' an accurate answer should be '[p]ossibly. Possibly not' (p. 75). Pawson demonstrates clearly the dangers of making unequivocal claims on the basis of under-researched, over-interpreted, neglectful data and narrow enquiry, and he counsels caution in making simplistic claims. As he says, there are reasons why we may not believe the claims made (p. 16), and evaluators worth their salt would do well to keep this to the fore in understanding how educational interventions unfold. This may come as unwelcome news to policy makers who seek unequivocal results.

Pawson notes that the search for 'what it is about an intervention that works for whom, in what circumstances, in what respects, over which duration' (2013, p. 167), i.e. to focus on the contexts, mechanism and outcomes of the intervention, is a complex, pragmatic and ongoing endeavour. In this he notes for evaluators the significance of context (pp. 36–8) and of attention to the personal and interpersonal dimensions of interventions (pp. 127–31, 139–46). Epistemological and substantive support for his approach finds voice in: (a) Popperian doubt, with its emphasis on conjectures, refutations and the tentative, conditional, falsifiable nature of scientific 'truths' (p. 99); and (b) Rossi's disconcerting 'iron law' (p. 12) which states that the value of any assessment of impact of a large-scale programme is typically zero (p. 12). This may be uncomfortable for both policy makers and evaluators, but that is the reality of many interventions.

In putting forward his case for a realist approach to evaluation, Pawson uses acronyms to identify key elements and foci of evaluation:

- CMOs: Contexts, Mechanisms and Outcomes (pp. 21–6) (redolent of Stufflebeam's (1967) Context, Input, Process and Product model of educational evaluation);

- VICTORE: Volitions, Implementation, Contexts, Time, Outcomes, Rivalry and Emergence (pp. 33–46);
- TARMATO: Theory, Abstraction, Reusable Conceptual Platforms, Model Building, Adjudication, Trust and Organized Scepticism (pp. 85–111).

Even though it may not be good news for policy makers, Pawson notes that many interventions typically work and then don't work. This, he avers, is due in part to context, situations, perspectives, participants, circumstances and affinity to decision makers' agendas. He notes that an evaluation is prone to distortion and misrepresentation of the multiple factors in, and interpretations of, an intervention phenomenon if: (a) it neglects understanding how the perceptions, actions, agency and circumstances of humans unfold in, and affect, an intervention; and (b) it is selective on whose interpretation of a project is adopted. This is salutatory advice for both evaluators and policy makers.

The effectiveness of interventions vary from school to school and from individual to individual (see also Cartwright and Hardie, 2012). Policy makers are attracted by 'what works'. However, Pawson (2013) shows that what works for some is unlikely to work for all, that it is massive over-determination rather than the intervention itself which frequently contributes to policy uptake, regardless of impact, and that, even though some policies are doomed to be implemented, to expect them to provide only gains rather than losses is simply naive. In other words, 'what works' is problematic. This being the case, evaluation seeks to identify the contingencies and conditions surrounding a decision or intervention, for example, if we want to do such-and-such then it would be better to adopt approach A and B, targeted at M and N, and to be aware of dangers of X and Y (cf. Pawson, 2013, p. 190).

5.3 Research, evaluation, politics and policy making

Evaluation and research are beset with issues of politics; they take place within a political environment, which might be a micro-environment (e.g. a single school) or a larger environment (e.g. a funding body, a research institute). MacDonald (1987) comments that the evaluator:

> is faced with competing interest groups, with divergent definitions of the situation and conflicting informational needs.… He has to decide which decision-makers he will serve, what information will be of most use, when it is needed and how it can be obtained.… The resolution of these issues commits the evaluator to a political stance, an attitude to the government of education. No such commitment is required of the researcher. He stands outside the political process, and values his detachment from it. For him the production of new knowledge and its social use are separated. The evaluator is embroiled in the action, built into a political process which concerns the distribution of power, i.e. the allocation of resources and the determination of goals, roles and tasks.… When evaluation data influences power relationships the evaluator is compelled to weight carefully the consequences of his task specification. … The researcher is free to select his questions, and to seek answers to them. The evaluator, on the other hand, must never fall into the error of answering questions which no one but he is asking.
>
> (MacDonald, 1987, p. 42)

Whether that holds as true at the present moment as when it was written is a moot point, as funded research often has a strong political motive, and institutional politics (e.g. internal funding decisions) may have a bearing on research. MacDonald argues that evaluation is an inherently political enterprise. His much-used threefold typology of evaluations as autocratic, bureaucratic and democratic is premised on a political reading of evaluation (see also Chelinsky and Mulhauser (1993), who refer to 'the inescapability of politics' (p. 54) in the world of evaluation).

The reality of politics often blurs distinctions between research and evaluation. Two principal causes of this blurring lie in the *funding* and the *politics* of both evaluation and research. For example, the view of research as uncontaminated by everyday life is naive and simplistic; Norris (1990, p. 99) argues that such an antiseptic view of research ignores the social context of educational research, some of which is located in the hierarchies of universities and research communities and the funding support provided for some but not all research projects by governments. His point has a pedigree that reaches back to Kuhn (1962), and is a comment on the politics of research funding and research utilization. For decades one can detect a huge rise in 'categorical' funding of projects, i.e. defined, *given* projects (often by government or research sponsors) for which bids have to be placed. This may seem unsurprising if one is discussing research grants from government bodies, which are typically deliberately policy-oriented, though one can also detect in projects which have been granted by non-governmental organizations a move towards sponsoring policy-oriented projects rather than the 'blue-skies' research mentioned

earlier, and the rise of the evidence-based movement in research and policy making draws this link ever tighter. Indeed Burgess (1993b) argues that 'researchers are little more than contract workers.… [R]esearch in education must become policy relevant.… [R]esearch must come closer to the requirement of practitioners' (p. 1), echoing Stenhouse's (1975) advocacy of the teacher-as-researcher and more recently in Pring's (2015) comments on the centrality of practitioner research in educational research. Burgess's view also points to the constraints under which research is undertaken; if it is not concerned with policy issues then research may not be funded, and research must have some impact on policy making.

The view of the tension between research, evaluation and politics is reinforced by several articles in the collection edited by Anderson and Biddle (1991) which show that research and politics go together uncomfortably because researchers have different agendas and longer timescales than politicians and try to address the complexity of situations, whereas politicians, anxious for short-term survival, want telescoped timescales, simple remedies and research that will be consonant with their political agendas. As James (1993) notes:

> the power of research-based evaluation to provide evidence on which rational decisions can be expected to be made is quite limited. Policy-makers will always find reasons to ignore, or be highly selective of, evaluation findings if the information does not support the particular political agenda operating at the time when decisions have to be made.
>
> (James, 1993, p. 135)

Her comments demonstrate a remarkable prescience as, if anything, the situation has become even more acute than when it was written.

Not only is *research* a political issue, but this extends to the use being made of *evaluation* studies. Whilst evaluations can provide useful data to inform decision making, as evaluation has become more politicized so its uses (or non-uses) have become more politicized. Indeed Norris (1990) provides examples of how politics frequently overrides evaluation or research evidence, and, despite the evidence-based movement, it is common to read how politicians introduce interventions in education on the basis of poor, scant or, indeed, no evidence of their efficacy. Gorard (2005, 2014), for example, demonstrates that 'academies' in the UK were doomed to succeed as there was no evidence to suggest that they were any better than the school which they replaced. This echoes James's earlier comment (1993) where she writes:

> The classic definition of the role of evaluation as providing information for decision makers … is a fiction if this is taken to mean that policy-makers who commission evaluations are expected to make rational decisions based on the best (valid and reliable) information available to them.
>
> (James, 1993, p. 119)

Where evaluations are commissioned and have heavily political implications, Stronach and Morris (1994) argue that the response to this is that evaluations become more 'conformative', possessing several characteristics:

1. short-term, taking project goals as given and supporting their realization;
2. ignoring the evaluation of longer-term learning outcomes, or anticipated economic/social consequences of the programme;
3. giving undue weight to the perceptions of programme participants who are responsible for the successful development and implementation of the programme; as a result, tending to 'over-report' change;
4. neglecting and 'under-reporting' the views of classroom practitioners and programme critics;
5. adopting an atheoretical approach, and generally regarding the aggregation of opinion as the determination of overall significance;
6. involving a tight contractual relationship with the programme sponsors that either disbars public reporting, or encourages self-censorship in order to protect future funding prospects;
7. undertaking various forms of implicit advocacy for the programme in its reporting style;
8. creating and reinforcing a professional schizophrenia in the research and evaluation community, whereby individuals come to hold divergent public and private opinions, or offer criticisms in general rather than in particular, or quietly develop 'academic' critiques which are at variance with their contractual evaluation activities, alternating between 'critical' and 'conformative' selves.

The points raised so far can apply to large-scale and small-scale projects. Hoyle (1986), for example, notes that evaluation data are used to bring resources into, or take resources out of, a department or faculty. In this respect the evaluator may have to choose carefully his or her affinities and allegiances (Barton, 2002), as the outcomes and consequences of the evaluation may call these into question. Barton writes that, although the evaluator may wish to remain passive and apolitical, in

reality this view is not shared by those who commission the evaluation or the reality of the situation, not least when the evaluation data are used in ways that distort the data or use them selectively to justify different options (p. 377).

The issue relates to both evaluations and research, as school-based research is often concerned more with finding out the most successful ways of organization, planning, teaching and assessment of a *given agenda* rather than *setting agendas* and following one's own research agendas. This is *problem solving* rather than *problem setting*. That evaluation and research are being drawn together by politics at both a macro- and micro-level is evidence of a continuing interventionism by politics into education, reinforcing the hegemony of the government in power.

Several points have been made so far:

- There is considerable overlap between evaluation and research;
- There are some conceptual differences between evaluation and research, though, in practice, there is considerable blurring of the edges of the differences between the two;
- The funding and control of research and research agendas often reflect the persuasions of political decision makers;
- Evaluative research has increased in response to categorical funding of research projects;
- The attention being given to, and utilization of, evaluation varies according to the consonance between the findings and their political attractiveness to political decision makers.

There is very considerable blurring of the edges between evaluation and research because of the political intrusion into, and use of, these two types of study. One response to this can be seen in Burgess's (1993a) view that a researcher needs to be able to meet the sponsor's requirements for evaluation whilst also generating research data (engaging the issues of the need to negotiate ownership of the data and intellectual property rights); for an example of this, see Garcia *et al*. (2014).

Research and politics are inextricably bound together. Researchers in education are advised to give serious consideration to the politics of their research enterprise and the ways in which politics can steer research (Hammersley, 2014). For example, one can detect a trend in educational research towards more evaluative research, where, for example, a researcher's task is to evaluate the effectiveness (often of the implementation) of given policies and projects. This is particularly true in the case of 'categorically funded' and commissioned research – research which is funded by policy makers (e.g. governments, fund-awarding bodies) under any number of different headings that those policy makers devise. On the one hand this is laudable, for it targets research directly towards policy (e.g. the 'what works' initiatives); on the other hand it is dangerous in that it enables others to set the research agenda. Research ceases to become open-ended, pure research and, instead, becomes the evaluation of *given* initiatives.

Evaluators may have the power to control the operation of the evaluation project and may influence the brief given, whilst the sponsor can only support but not control the independence of the evaluator. The issue of sponsoring research reaches beyond simply commissioning research towards the dissemination (or not) of research: who will receive or have access to the findings and how the findings will be used and reported. This, in turn, raises the fundamental issue of who owns and controls data, and who controls the release of research findings. Unfavourable reports might be withheld for a time, suppressed or selectively released. In other words, research can be brought into the service of wider educational purposes, for example, the politics of a local education authority, or indeed the politics of government agencies.

Though research and politics intertwine, the relationships between educational research, politics and policy making are complex because research designs strive to address a complex social reality (Anderson and Biddle, 1991). A piece of research does not feed simplistically or directly into a specific piece of policy making. Rather, research generates a range of different types of knowledge: concepts, propositions, explanations, theories, strategies, evidence and methodologies. These feed subtly and often indirectly into the decision-making process, providing, for example, direct inputs, general guidance, a scientific gloss, orienting perspectives, generalizations and new insights.

The degree of influence exerted by research depends on careful dissemination; too little and its message is ignored, too much and data overload confounds decision makers and makes them cynical – the syndrome of the boy who cried wolf (Knott and Wildavsky, 1991). Hence researchers must give care to utilization by policy makers (Weiss, 1991a), reduce jargon, provide summaries and improve links between the two cultures of researchers and policy makers (Cook, 1991) and, further, to the educational community. Researchers must cultivate ways of influencing policy, particularly when policy makers can simply ignore research findings, commission their own research (Cohen and Garet,

1991) or underfund research into social problems (Coleman, 1991; Thomas, 1991). Researchers must recognize their links with the power groups who decide policy. Research utilization takes many forms depending on its location in the process of policy making, for example, in research and development, problem solving, interactive and tactical models (Weiss, 1991b).

The impact of research on policy making depends on its degree of consonance with the political agendas of governments (Thomas, 1991) and policy makers anxious for their own political survival (Cook, 1991) and the promotion of their social programmes. Research is used if it is politically acceptable. That the impact of research on policy is intensely and inescapably political is a truism. Research too easily becomes simply an 'affirmatory text' which 'exonerates the system' (Wineburg, 1991) and is used by those who seek to hear in it only echoes of their own voices and wishes (Kogan and Atkin, 1991).

There is a significant tension between researchers and policy makers. The two parties have different, and often conflicting, interests, agendas, audiences, timescales, terminology and concern for topicality (Levin, 1991). These have huge implications for research styles. Policy makers anxious for the quick fix of superficial facts seek unequivocal data, short-term solutions and simple, clear remedies for complex and generalized social problems (Cartwright, 1991; Cook, 1991; Radford, 2008, p. 506; Cartwright and Hardie, 2012): the Simple Impact model (Biddle and Anderson, 1991; Weiss, 1991a, 1991b). Moreover, policy makers often find much research too uncertain in its effects (Kerlinger, 1991; Cohen and Garet, 1991), too unspecific and too complex in its designs, and of limited applicability (Finn, 1991). This, reply the researchers, misrepresents the nature of their work (Shavelson and Berliner, 1991) and belies the complex reality which they are trying to investigate (Blalock, 1991). Capturing social complexity and serving political utility can run counter to each other. As Radford (2008, p. 506) remarks, the work of researchers is often driven by objectivity and independence from, or disinterestedness in, ideology, whereas policy makers are driven by interests, ideologies and values.

The issue of the connection between research and politics – power and decision making – is complex. On another dimension, the notion that research is inherently a political act because it is part of the political processes of society has not been lost on researchers (cf. Hammersley, 2011, 2014), and this harks back to Chapter 3 in its discussion of value-neutrality and partisan research. Researchers cannot be blind to politics, just as politicians and decision makers should not be blind to research evidence and evaluation. As evaluation and research draw ever closer together it is prudent for researchers to reflect on Becker's (1967) rallying call for researchers to consider carefully 'whose side are we on'.

The companion website to the book provides PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: www.routledge.com/cw/cohen.

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

The search for causation

CHAPTER 6

This chapter introduces key issues in understanding causation in educational research. These include:

- causes and conditions
- causal inference and probabilistic causation
- causation, explanation, prediction and correlation
- causal over-determination
- the timing and scope of the cause and the effect
- causal direction, directness and indirectness
- establishing causation
- the role of action narratives in causation
- researching causes and effects
- researching the effects of causes
- researching the causes of effects

6.1 Introduction

Our brains seemed to be hard-wired to think causally, but, for educational researchers, tracing and using causality is challenging. Working with cause and effect, researchers must address the importance of being able to use educational research findings in the 'real world', that is, the world in which we have not isolated and controlled out a swathe of pertinent factors, or made so many assumptions and *ceteris paribus* get-out clauses (Kincaid, 2009) as to render the research of little or no value, or found effects which are wonderful in the sanitized, artificial world of the laboratory but useless in the 'real world' outside it (e.g. Cartwright and Hardie, 2012).

There is a wide vocabulary of causality in research, with many relevant words and phrases: for example, 'are caused by', 'influence' (verb and noun), 'attributed to', 'depend on', 'impact' (verb and noun), 'effect' (verb and noun), 'direction of the relationship', 'positive influence', 'positive impact', 'result' (verb and noun), 'mediation effect', 'effect of', 'due to', 'condition' (verb), 'leading to', 'consequences of', 'because of', 'affect', 'reason for', 'to force', 'driven', 'lead to' etc. Though the vocabulary is varied, the issue of demonstrating causality remains a challenge to researchers.

Educationists and social scientists are concerned not only for 'what works' but 'why', 'how', 'for whom' and 'under what conditions and circumstances'. They want to predict what will happen if such-and-such an intervention is introduced, and how and why it will produce a particular effect. This points us to an important feature of educational research, which is to look for causation: what are the effects of causes and what are the causes of effects? This is not a straightforward enterprise, not least because causation is often not observable but can only be inferred, and it is highly unlikely that indisputable causality is ever completely discoverable in the social sciences. At best probabilistic causation offers a more fitting characterization of causation in educational research. Causation is often considered to be the 'holy grail' of educational research, and this chapter introduces some key considerations in investigating causation.

6.2 Causes and conditions

Novice researchers are faced with many questions concerning causation in their research, for example:

- whether the research is seeking to establish causation, and if so, why;
- deciding when causation is demonstrated, recognizing that causation is never 100 per cent certain;
- deciding what constitutes a cause and what constitutes an effect;
- deciding what constitutes evidence of the cause and evidence of the effect;
- deciding the kind of research and the methodology of research needed if causation is to be investigated;
- deciding whether the research is investigating the cause of an effect, the effect of a cause, or both.

To infer simple, deterministic or regular causation may be to misread many situations, excepting, perhaps, those where massive single causation is clear. It may be more useful for the researcher to consider causal processes rather than single events (Salmon, 1998), not least because there is often more than a single cause at work in any effect and there may be more than one effect from a single cause. Indeed, the researcher has to

distinguish between causes, reasons, motives, determination and entailment, and whilst these might all exert causal force in some circumstances or enable us to make causal explanations or predictions, in other circumstances they do not.

What, then, makes a cause a cause, and an effect an effect? How do we know? Though we can say that causation takes places in a temporal sequence – the cause precedes the effect, and with temporal succession (Hume's criterion of 'priority'; Hume, 1955; Norton and Norton, 2000), this does not help the researcher very much.

One distinguishing indication that causation is taking place or has taken place is the presence of counterfactuals (Mackie, 1993), i.e. the determination that the absence of X (the supposed cause) would have led to the absence of Y (the effect); 'if X had not happened then Y would not have happened'. If we are seeking to establish that such-and-such is a contributing cause (X) of an effect (Y) we ask ourselves whether, if that supposed cause had not been present, then would the effect have occurred or been what it actually was; if the answer is 'no' then we can suppose that X is a true cause. For example, if there had been no ice on a path then I would not have fallen over and broken my arm. So the presence of ice must have been a contributing cause of the effect, one of many causes (e.g. my poor sense of balance, my poor eyesight in not seeing the ice, the ambient darkness, wearing slippery-soled shoes, brittle bones because of age etc.).

The counterfactual argument is persuasive, but problematical: how do we know, for example, what the outcome would have been if there had been no patch of ice on the path where I was walking? Can we predict with sufficient certainty to attribute counterfactual causality here? How can we prove that the effect would *not* have happened if a particular cause had not been present? How do we know that I would or would not have slipped and fallen if the ice had not been present? In true experiments this is addressed by having a control group: the control group is supposed to indicate what would have happened if the intervention had not occurred. The problem is that much research is not experimental.

If it were only the presence of ice that caused me to fall and break my arm, then this would be a very simple indication of causality; the problem is that the presence of ice in this instance is perhaps not a sufficient cause – had my balance been good, my eyesight good, the ambient light good, and if my shoes had had good grips on their soles and my bones were less brittle, then I would not have fallen and broken my arm.

The difficulty here also is to establish the role (if any) and relative strength of the causes in a multi-causal situation, i.e. the several conditions that, themselves, contribute to the accident. The presence of those causes that are included affects their relative strengths in a specific context, and the absence of some of these causes in the same context may raise or lower the relative strengths of others.

The example of falling on the ice also indicates an important feature of causation: causes cannot be taken in isolation, they may need to be taken together (compound causes, i.e. they only exert causative force when acting in concert), and there may be interaction effects between them. On its own, the patch of ice might not have caused my fall and broken arm; it was perhaps neither sufficient nor necessary, as I could have fallen and broken my arm anyway because of my slippery shoes and poor balance. On its own, my poor balance did not cause me to fall and break my arm. On its own, the darkness did not cause me to fall and break my arm. On their own, my slippery soles did not cause me to fall and break my arm. On their own, my brittle bones did not cause me to fall and break my arm. But put all these together and we have sufficient conditions to cause the accident. For the researcher, looking for individual causes in a contextualized situation may be futile.

In understanding the causes of effects, one has to understand the circumstances and conditions in which the two independent factors – the cause and the effect – are located and linked (the link is contingent rather than analytic). Discovering the circumstances – conditions – in which one variable causes an effect on another is vital in understanding causation, for it is the specific combination of necessary and/or sufficient conditions that may produce an effect. Causes of effects work in specific circumstances and situations, and account has to be taken of these circumstances and conditions.

For the researcher, the difficulty in unravelling the effects of causes and the causes of effects is heightened by the fact that causes may be indirect rather than direct (cause A causes effect B, and effect B causes effect C) or that they may only become a cause in the presence of other factors (I may fall over on ice and not break my arm if I am young and land well, but, as an older person with more brittle bones, I may land awkwardly and break my arm – the fall is not a sufficient condition or cause of my broken arm).

6.3 Causal inference and probabilistic causation

Identifying and understanding causation may be problematic for researchers, as effects may not be direct linear functions of causes, and because there may be few, many, increasing, reducing, unpredictable, i.e.

non-linear effects of causes (see Chapter 1 on complexity theory). A small cause can bring about a large or irregular effect (or, indeed, no effect, in the presence of other factors); a large cause may bring about a small or irregular effect (or, again, no effect, in the presence of other factors). Causation is often an inductive and empirical matter rather than a logical, deductive matter, and, indeed, it is often unclear what constitutes a cause and what constitutes an effect as these are often umbrella terms, under which are sub-causes and sub-effects, causal processes, causal chains, causal webs and causal links bringing several factors together both at a particular point in time (the moment of falling, in the example above) and in a temporal sequence.

Further, there is an asymmetry at work in causation effect – a cause can produce an effect but not vice versa: being young, good-looking and female may help me to pass my driving test if I am in the presence of a leering male examiner, but passing my driving test does not cause me to be young, good-looking and female.

It is often dangerous to say that such-and-such is *definitely* the cause of something, or that such-and-such is *definitely* the effect of something. Causation in the human sciences is much more tentative, and may be probabilistic rather than deterministic. Hume's (2000) own rules for causation are:

- contiguity (of space and time) (the cause is contiguous with the effect);
- priority/succession (the cause precedes the effect);
- constant conjunction (the coupling of one event and its successor are found to recur repeatedly);
- necessary connection (which is learned from experience, habit and custom rather than from deductive, logical, necessary proof).

One can detect correlation in Hume's ideas rather than actual causation. He argues that causation is inferred, inductively, by humans rather than being an objective matter. To try to gain some purchase on causality, Mill (2006) sets out five main approaches to establishing causality, and these are outlined below.

Mill's method of *agreement*: Let us say that in different regions of a country there are several combined educational reforms taking place, designed to increase student mobility (Table 6.1): (a) increased educational financing; (b) curriculum reforms; (c) providing more places in vocational training; and (d) introducing National Qualifications Frameworks (NQFs). We wish to see which of these factors is causing increased student mobility.

Mill's method of agreement here states that in all the cases where the effect occurs ('increased student mobility'), if there is only one factor common to all the cases then that factor is the cause. Here only one column ('Introducing the NQF') has all four factors present. Is it safe to conclude that, *ceteris paribus*, the NQF is the cause of the increased student mobility? Perhaps. If these were the only relevant factors in the situation then the conclusion might be safe, but, of course, it is not; the real situation includes far more factors. Mill's method is an over-simplification in the empirical world.

Mill's method of *difference*: Let us say that different regions of a country have several reforms taking place, designed to increase student mobility (Table 6.2). One region did not have an NQF, and this is the only factor where there is no increased student mobility. Is it safe to conclude that the NQF is the cause of the increased student mobility? Perhaps. As before, if these were the only relevant factors in the situation then the conclusion might be safe, but in the 'real' world of multiple factors, it is not.

Mill's method of *agreement and difference*: This applies both the preceding methods (Table 6.3) together. Here the method suggests that the NQF may be the cause of increased student mobility:

1 It could not be 'increased educational financing', as it is present where there is no 'increased student mobility' in region 1.

TABLE 6.1 MILL'S METHOD OF AGREEMENT

Region	*Increased educational financing*	*Curriculum reforms*	*More places for vocational training*	*Introducing the NQF*	*Increased student mobility*
Region 1	✓	✓	✓	✓	✓
Region 2	No	✓	No	✓	✓
Region 3	✓	No	No	✓	✓
Region 4	No	No	✓	✓	✓

TABLE 6.2 MILL'S METHOD OF DIFFERENCE

Region	*Increased educational financing*	*Curriculum reforms*	*More places for vocational training*	*Introducing the NQF*	*Increased student mobility*
Region 1	✓	✓	✓	No	No
Region 2	✓	✓	✓	✓	✓
Region 3	✓	✓	✓	✓	✓
Region 4	✓	✓	✓	✓	✓

TABLE 6.3 MILL'S METHOD OF AGREEMENT AND DIFFERENCE

Region	*Increased educational financing*	*Curriculum reforms*	*More places for vocational training*	*Introducing the NQF*	*Increased student mobility*
Region 1	✓	No	✓	No	No
Region 2	✓	✓	No	✓	✓
Region 3	✓	No	✓	✓	✓
Region 4	✓	✓	✓	✓	✓

2 It could not be 'curriculum reforms', as it is absent where there is 'increased student mobility'.
3 It could not be 'more places for vocational training', as it is absent where there is 'increased student mobility'.
4 This leaves NQF, which is the only item which is remaining when (1) to (3) above are taken into account.

As with the two previous cases, is it safe to conclude that the NQF is the cause of the increased student mobility? Perhaps. Again, if these were the only relevant factors in the situation then the conclusion might be safe, but in the 'real' world, it is not.

Mill's method of *concomitant variation*: If, across several factors, one finds that the property (e.g. the amount) of variation in Factor (A) is similar to, or the same as, the amount of variation in the effect (Factor (B)), whilst such common variation is not demonstrated in other independent variables, then it may be reasonable to infer that Factor (A) is the cause of Factor (B) (Table 6.4, matching the row entry for each cause with the effects on the same row). Table 6.4 indicates that the only concomitant variation between the independent and dependent variable ('increased student mobility') is for the variable 'introducing the NQF', suggesting that it is the NQF which might be causing the effect. Is it safe to conclude that the NQF is the cause of the increased student mobility? Perhaps. Again, if these were the only relevant factors in the situation then the conclusion might be safe, but in the 'real' world, it is not.

Mill's method of *residues*: If one is able to remove (e.g. control out) all the factors but one that may be causing all the effects but one, then the remaining factor is the cause of the remaining effect (Table 6.5).

Here, we have a range of possible causes and effects; we see that all the factors except 'introducing the NQF' are causing all the effects (those in the top row of Table 6.5) except one ('increased student mobility'). Hence we hold that it is 'introducing the NQF' which is the cause of 'increased student mobility'. Is it safe to conclude that the NQF is the cause of the increased student mobility? Perhaps. As with all the previous four methods from Mill, if these were the only relevant factors in the situation then the conclusion might be safe, but in the 'real' world, it is not.

Mill's methods are only as powerful as the factors included, and, in their search for the single cause, may overlook the interplay of myriad causes in producing the effect and, indeed, assume a deterministic rather than probabilistic view of causation (Kincaid, 2009). Further, Mill's approach to establishing causality in this instance operates at a single country level. Imagine,

TABLE 6.4 MILL'S METHOD OF CONCOMITANT VARIATION

Causes			*Effect*		
Increased educational financing			*Increasing student mobility*		
A little	Moderate amount	A lot	A little	Moderate amount	A lot
✓				✓	
		✓		✓	
		✓		✓	
Curriculum reforms			*Increasing student mobility*		
A little	Moderate amount	A lot	A little	Moderate amount	A lot
✓					✓
	✓		✓		
		✓		✓	
More places for vocational training			*Increasing student mobility*		
A little	Moderate amount	A lot	A little	Moderate amount	A lot
✓				✓	
	✓				✓
		✓	✓		
Introducing the NQF			*Increasing student mobility*		
A little	Moderate amount	A lot	A little	Moderate amount	A lot
✓			✓		
	✓			✓	
		✓			✓

TABLE 6.5 MILL'S METHOD OF RESIDUES

Causes	*Effects*			
	More students in higher education	*Greater vocational relevance*	*Clearer progression in qualifications*	*Increased student mobility*
Increased educational financing	✓	✓	✓	
Curriculum reforms	✓	✓	✓	
More places for vocational training	✓	✓	✓	
Introducing the NQF	✓	✓	✓	✓

then, the additional complexity where his approach is applied to more than one country at a time; the problem expands exponentially in trying to detect the causal links between one factor and its putative effects. Nevertheless Mill's view has substantially informed more recent enterprises in working with causation, for example in the work of Ragin (1987, 2008).

Whilst statistical modellers may argue that it is possible to operate controls and to utilize structural equation modelling, these necessarily simplify the complex, dynamic, changing scenario of the relationship between the NQF, its instigator, its context and the multiple outcomes operating in a situation. It is impossible to isolate and control variables in a situation (cf. Cartwright

and Hardie, 2012); indeed the situation may rely on the dynamic interplay of these variables. This frustrates any easy attempts to utilize Mill's (2006) five main approaches to establishing causality.

Methods for establishing whether an effect is truly the result of a cause are beset with problems. Even if we observe outcomes, we cannot conclude with any certainty that these are unequivocally caused by the NQF in the example here. Even though probabilistic causality may replace deterministic causality, this does not attenuate the problem of deciding what is legitimate and illegitimate inference.

The inferential, conjectural and probabilistic nature of much causation in educational research (rather than being absolute, deductive and deterministic), coupled with the fact that causation is frequently unobservable, renders the study of causation challenging for educational researchers. Indeed there is a danger in isolating and focusing on singular causes separately from other contributing causes, contexts and conditions, and it is perhaps more fitting to regard causes as processes over time rather than single events. Further, in an interconnected world of multiple causes and causal nets, conditions and interactions may provide better accounts of causation than linear determinism (Morrison, 2012).

In unravelling causes and effects, the researcher is faced with the task of identifying what actually constitutes a cause and what constitutes an effect. The contexts and conditions of an event are as important as the trigger of an effect, and may be contributing causes. In the example earlier, my falling and breaking my arm was precipitated – triggered – by the ice on the path, but, without the presence of other contributing factors I might not have fallen and I might not have broken my arm. The trigger of the effect may not be its sole cause but only the last cause in a causal chain, sequence of events, series or network of conditions before the effect occurs, even though causes often raise the likelihood of their effects rather than guaranteeing them (Mellor, 1995, pp. 69–70). Indeed, whilst probability often concerns identifying *likelihood*, the strongest probability is not always the same as the strongest causation. I might think that putting pressure on a child to succeed has the strongest possibility of causing her success, but the actual cause might lie elsewhere, for example, the teacher might be very effective, the students might be highly motivated or the examination might be very easy.

6.4 Causation, explanation, prediction and correlation

The demonstration of causation is difficult. Causation is not the same as explanation (Clogg and Haritou, 1997, p. 106; Salmon, 1998, pp. 5–8) (e.g. an explanation may be wrong, or it may be giving the meaning of something, or it may be indicating how to do something). Nor is causation the same as giving a reason. For example, I might take a day off from work, giving the reason that I am sick, but the real reason may be simply that I am lazy or want to go shopping.

Nor is causation the same as prediction. Just because I observe something happening once does not mean I can predict that it will happen again (the problem of induction, see Chapter 1), as the conditions could be different, or, indeed, even if the conditions were very similar (as chaos theory tells us). I might be able to predict something even though my prediction is based on the wrong identification of causes, for example, I can predict that there will be a storm because I have observed the barometric reading falling, but the fall in the barometric reading does not cause the storm. Formally put, the two variables – the barometric reading and the storm – are 'screened off', separated and kept apart from each other (Reichenbach, 1956; Salmon, 1998). They have correlational but no causal relationships to each other, and are both caused by a third factor – the drop in air pressure (see Figure 6.1).

I might predict that a person's hands might be large if she has large feet, but having large hands does not cause her to have large feet – the cause might lie in a genetic predisposition to both. It is one thing to say that a change in one variable (A) is associated with a change in another variable (B); it is an entirely different thing to say that a change in one variable (A) *brings about* a change in another variable (B); and it is an entirely different thing again to say that a change in one variable (B) is brought about by a change in another variable (A), i.e. that it *is caused* by that change in variable A.

In attributing genuine causation, it is useful to 'screen off' unrelated dependent variables from the variables that are directly relevant to the situation being researched, i.e. to ensure that the effect of one variable is removed from the equation, to discount that variable or to control for the effects of other variables, for

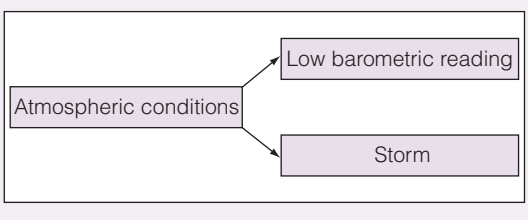


FIGURE 6.1 Two unrelated factors caused by a third factor

example, the presence of a third variable or several variables, by partial correlations. It is also important in screening off to ensure that one variable is not deemed to have an influence on another when, in fact, this is not the case. This presumes that it is actually possible to identify which factors to screen off from which. Pearl (2009, pp. 423–7) indicates how this can be approached; in the case of multiple causality (or in the cases of over-determination, discussed below), this may not be possible.

Screening off requires the ability to separate out causes, and this may be difficult to the point of impossibility. However, in seeking to establish genuine causation, the researcher must consider controlling for the effects of additional variables, be they prior/exogenous variables or intervening/endogenous variables, as these might exert a non-causal influence on the dependent variables.

In conducting research that seeks to attribute causation, it is important to control for the effects of variables, i.e. to hold them constant so that fair attribution of causality and the weight of causal variables can be assessed (though relative weights of causes are, strictly speaking, superfluous in discussing causation; they are questionable indicators of causation). Further, identifying the relative strengths of causes depends on the presence or absence of other causes. For example, in looking at examination success (the effect), if my research confines itself to looking at the relative strength of causes A (hours of study), B (IQ) and C (motivation) in producing the effect, I might find that C (motivation) is the strongest of the three causes. However, if I were to add a new variable (D) (an outstanding teacher helping the student), then it may be that D is the overriding cause and that A, B and C are of equally low strength, or that A becomes the second strongest factor.

Statistical tools such as crosstabulations, correlation and partial correlation, regression and multiple regression, and structural equation models (see Chapters 40–43) can be used to assist here in the analysis of causation, though it is often difficult to control direct, indirect, antecedent, intervening and combined influences of variables on outcomes (though statistical tools and graphical methods can assist here; Pearl, 2009, pp. 423–7).

In considering the control of variables, let us examine, for example, the subject choices of secondary school male and female students (Table 6.6).

Here we can see overwhelmingly that males choose physics far more than females, and females choose biology far more than males. The researcher wishes to know if the allocation of certain teachers to teach the secondary school science subjects affects the students' choice (i.e. whether it is the subject or the teacher, or some combination of these, that is causing the students to choose the subjects that they choose). The researcher introduces the third variable of the 'teacher' as a control variable, with two values: Teacher A and Teacher B, and then partitions the data for males and females according to either Teacher A or Teacher B (see Table 6.7).

TABLE 6.6 SCIENCE CHOICES OF SECONDARY SCHOOL MALES AND FEMALES

	Male	*Female*	*Total*
Preference for physics	175 (55.1%)	87 (27.9%)	262 (41.6% of total)
Preference for biology	143 (44.9%)	225 (72.1%)	368 (58.4% of total)
Column total	318 (100%)	312 (100%)	630
Percentage of total	50.5%	49.5%	100%

TABLE 6.7 SCIENCE CHOICES OF MALE AND FEMALE SECONDARY STUDENTS WITH TEACHER A OR B

	Males with Teacher A	*Females with Teacher A*	*Males with Teacher B*	*Females with Teacher B*	*Total*
Preference for physics	86 (55.8%)	44 (28.4%)	89 (54.3%)	43 (27.4%)	262 (62.5% of total)
Preference for biology	68 (44.2%)	111 (71.6%)	75 (45.7%)	114 (72.6%)	368 (37.5% of total)
Column total	154 (100%)	155 (100%)	164 (100%)	157 (100%)	630
Percentage of total	24.4%	24.6%	26.1%	24.9%	100%

When the data are partitioned by teacher (Teacher A and Teacher B) the researcher notes that the percentages in each of the partial tables (one part of the table for Teacher A and the other part of the table for Teacher B) in Table 6.7 are very similar to the original percentages of the root table (Table 6.6). She concludes that whether Teacher A or Teacher B is teaching the class makes no appreciable difference to the choices made by the students. The percentages in the new table (Table 6.7) replicate very closely those in the original (Table 6.6). The researcher concludes that the teacher involved is exerting no causal influence on the choice of subjects by the secondary school males and females.

However, let us imagine that the partial tables had yielded different data (Table 6.8). This time the results of the choices made by males and females who are with Teacher A and Teacher B are very different. The percentages in the new table (Table 6.8) are very different from those in the original (Table 6.6). This suggests to the researcher that, in this instance, the teacher of the class in question is making a causal difference to the choices of science subject made by the students.

However, this only tells us the 'what' of causation, or, to be more precise, it only gives us an indication of association and possible causation: it appears that the teacher makes no difference in Table 6.7 but *does* make a difference in Table 6.8. How this becomes a *causal* matter is another question altogether: how does the teacher *actually* affect the males' or females' choices of which science subject to follow. For example is it that: (a) Teacher A is male and Teacher B is female, and students tend to prefer to be with teachers of their own gender; (b) Teacher A has a better reputation than Teacher B for helping students to pass public examinations with high grades, and students are anxious to do well; (c) Teacher A is more sympathetic than Teacher B, so that students can relate more easily to Teacher A, and so they choose Teacher A; (d) Teacher A has a better sense of humour than Teacher B, and students prefer a good-humoured teacher; (e) Teacher A explains matters more clearly than Teacher B, and students prefer clear explanations, and so on. The point here is that, though one can deduce certain points from contingency tables and partial tables, they may not actually indicate causality. The same principle for holding variables constant, this time in correlational research, is discussed in Chapter 40.

In establishing causation, it is important to separate covariance and correlation between two unrelated and non-interacting dependent variables due to a common cause from the interaction of dependent variables due to the presence of a common cause (the examples of the barometer and the storm earlier).

6.5 Causal over-determination

It is rare to find a single cause of a single effect. It more often the case that there are several causes at work in a single situation and that these produce a multiplicity of effects (see the discussion of evaluation in Chapter 5, and the work of Pawson (2013) in that chapter). For example, why are so many young children well-behaved at school, when nobody has explicitly taught them the hidden curriculum (Jackson, 1968) of rules, regulations, taking turns, sharing, being quiet, knowing that the teacher is in charge and has all the power, putting up with delay, denial and only being one out of many children who has to gain the teacher's attention? One answer is over-determination: many events, both separately and in combination, lead to the same outcome – the young child must do as she is told, and having a nice time at school depends on how effectively she learns these rules and abides by them. Many causes; same effect: good behaviour.

Causal over-determination is 'where a particular effect is the outcome of more than one cause, each of which, in itself, would have been sufficient to have produced the effect' (Morrison, 2009, p. 51). A familiar example is the issue of which bullet can be said to have killed a man, which causes his death (Horwich, 1993),

TABLE 6.8 FURTHER SCIENCE CHOICES OF MALE AND FEMALE SECONDARY STUDENTS WITH TEACHER A OR B

	Boys with Teacher A	*Girls with Teacher A*	*Boys with Teacher B*	*Girls with Teacher B*	*Total*
Preference for physics	133 (56.8%)	55 (23.5%)	39 (46.4%)	35 (44.8%)	262 (62.5% of total)
Preference for biology	101 (43.2%)	179 (76.5%)	45 (53.6%)	43 (55.1%)	368 (37.5% of total)
Column total	234 (100%)	234 (100%)	84 (100%)	78 (100%)	630
Percentage of total	37.1%	37.1%	13.3%	12.5%	100%

if two bullets simultaneously strike a man's head. Either one bullet *or* the other caused the death (cf. Mellor, 1995, p. 102). Let us say that, in a study of homework and its effect on mathematics performance, a rise in homework might produce a rise in students' mathematics performance. However, this is not all: there may have been tremendous parental pressure on the child to do well in mathematics, or the student might have been promised a vast sum of money if her mathematics performance increased, or the school might have exerted huge pressure on the student to succeed, or the offer of a university place was contingent on a high mathematics score. The rise in mathematics performance may not have required all of the factors to have been present in order to bring about the effect; any one of them could have produced the effect. The effect is 'over-determined'. One effect may have one or several causes. Whilst this is commonplace, it is important to note this in order to refute claims frequently made by protagonists of such-and-such an intervention in education that it alone improves performance; if only it were that simple!

6.6 The timing and scope of the cause and the effect

Turn back to the earlier example of my falling on the ice and breaking my arm. Maybe I had a weakness in my arm from an injury many years before, and maybe when I injured my arm years before I could not have predicted that, many years later, I would have fallen on ice and broken my arm. The issue is not idle for researchers, for it requires them to consider, in terms of temporality, what are relevant causes and what to include and exclude from studies of causation, how far back in time to go in establishing causes and how far forward in time to go in establishing effects.

Just as the timing of causes may be unclear, so the timing of the effects of a cause may be unclear. Effects may be short-term only, delayed, instantaneous, immediate, cumulative and long-term; indeed the full effects of a cause may not be revealed in a single instance, as an effect may be a covering term for many effects that emerge over time (e.g. the onset and presenting of cancer has several stages; cancer is not a single event at one point in time). Temporality and causation are intimately connected but separate.

The examples above also indicate that terms such as 'cause' and 'effect' are, in many cases, shorthand for many sub-causes, sub-processes and sub-effects. Further, causes and effects may only reveal themselves over time, and, indeed, it may be difficult to indicate when a cause begins (which cigarette brought about the onset of cancer, or when did smoking first bring about the onset of cancer) or ends, and when an effect begins (e.g. I may continue smoking even after the early onset of lung cancer). I might hate studying mathematics at school but find it very attractive twenty years later; had my interest in mathematics been post-tested immediately I left school, the result would have been lower than if I had been tested twenty years later.

Where a cause begins and ends, where an effect begins and ends, when and how causes and effects should be measured, evaluated, ascertained and assessed, are often open questions, requiring educational researchers to clarify and justify their decisions on timings in isolating and investigating causes and effects. Quantitative data may be useful for identifying the 'what' of causation – what causes an effect – but qualitative data are pre-eminently useful for identifying the 'how' of causation – how causation actually works, the causal processes at work.

Consider, too, the reason for the ice patch being present on the path in the earlier example, and my being on the ice on the day in question. Maybe the local government services had not properly cleared the path of ice on that day, or maybe, as an ailing pensioner, I would normally be accompanied by a carer or an assistant whenever I went out, but on that day the person failed to turn up, so I was forced to go out on my own. Again, the issue is not idle for researchers, for it requires them to consider how widely or narrowly to cast their net in terms of looking for causes (how far out and how far in). In determining what are relevant causes, the researcher has to decide what to include and exclude from studies of causation, for example, from the psychological to the social, from the micro to the macro, and to decide the direction and combination of such causes.

The determination of a cause involves decisions on how far back to go in a temporal causal chain or network of events, and how wide or narrow to go in the causal space (how many conditions and circumstances contribute to the causation at work in a given situation). It may be difficult, if not impossible, to identify and include all the causal antecedents in a piece of research. Here the concepts of *necessary* and *sufficient* conditions are raised, as is the importance of identifying the causal trigger in a situation (the last cause in a causal chain or a linkage of several conditions). The striking of a match might cause it to flare, but that is not the only factor to be taken into account. Whether it flares depends on the abrasiveness of the striking surface, the materials used in the match, the dryness of the materials, the strength of the strike, the duration of the strike, the presence of sufficient oxygen in the atmosphere, and so on.

There may be an infinite number of causes and effects, depending on how far back one goes in time and how wide one goes in terms of contexts. This presents a problem of where to establish the 'cut-off' point in identifying causes of an effect. Whilst this may be addressed through the identification of necessary and sufficient conditions (Mackie, 1993) or the screening off of some 'ancestors' (antecedents) (Pearl, 2009), in fact this does little to attenuate the problem in social sciences, as not only is it problematic to identify what qualify as necessary or sufficient conditions, but these will vary from context to context, and even though there may be regularities of cause and effect from context to context, there are also differences from context to context.

The issue to be faced by researchers here is one of 'boundary conditions' and 'circumscription' (Pearl, 2009, p. 420): which factors we include or exclude can affect our judgements of causality. If, in a study of student performance, I only look at teacher behaviour and its influence on student performance then I might be led to believe that teacher behaviour is the cause of student performance, whereas if I only look at student motivation and its influence on student performance then I might be led to believe that student motivation is the cause of student performance. Researchers rarely, if ever, include the universe of conditions, but rather only a selection from that universe, and this might distort the judgements made about causation or where to look for causation. Whilst it may not be possible to identify the universe of conditions, the researcher has to be aware of the dangers of circularity, i.e. I am only interested in effect Y, so I only look at possible cause X, and then I find, unsurprisingly, that X is the cause of Y, simply because I have not considered alternatives. It is important to identify and justify the inclusion and exclusion of variables in researching causation, to select the field of focus sufficiently widely and to consider possible alternative explanations of cause and effect.

6.7 Causal direction, directness and indirectness

The problem of identifying causes and effects is further compounded by consideration of direct and indirect causes and effects and causal directions. Many models of causality often make too great a claim for unidirectionality rather than, for example, multi-directionality and mutual-directionality, or overlook clusters of causes that act together in multiple directions (Morrison, 2012).

Whilst the research may wish to identify the cause A that brings about the effect B, in practice this is seldom the case, as between A and B might be a huge number of intervening, prior or additional variables and processes operating, both exogenous and endogenous.

An exogenous variable is one whose values are determined outside the model (e.g. a structural equation model or a causal model) in which that variable is being used, or which is considered not to be caused by another variable in the model, or which is extraneous to the model.

An endogenous variable is one whose values or variations are explained by other variables within the model, or which is caused by one or more variables within the model. It is important to identify which causes mediate, and are mediated by, other causes. One also has to consider the role of moderators and mediator variables (the former influences the strength of a relationship between two variables, and the latter explains the relationship between two variables). How the researcher does this takes many forms, from theoretical modelling and testing of the model with data, to eliciting from participants what are the causes.

Whilst causation is not straightforward to demonstrate, this is not to suggest that establishing causation should not be attempted. There are regularities, likelihoods (probabilities) based on experience and previous research, and similarities between situations and people. Indeed the similarities may be stronger than the differences. This suggests that establishing probabilistic causation or inferring causation, whilst complex and daunting, may be possible for the researcher.

The problem for the researcher is to decide which variables to include, as the identification and inclusion/exclusion of relevant variables in determining causation is a major difficulty in research. Causes, like effects, might often be better regarded in conjunction with other causes, circumstances and conditions rather than in isolation (Morrison, 2009, 2012). Contextuality – the conditions in which the cause and effect take place – and the careful identification and inclusion of all relevant causes, are key factors in identifying causation.

6.8 Establishing causation

It is not easy to establish causation. For example, causation may be present but unobserved and indeed unobservable, particularly in the presence of stronger causes or impeding factors. I might take medication for a headache but the headache becomes worse; this is not to say that the medication has not worked, as the headache might have become even stronger without the medication. The effects of some causes may be masked by the presence or strength of others, but nonetheless causation may be occurring.

Morrison (2009, p. 45) gives an example where, in the case of the causal relationship between smoking (A), heart disease (B) and exercise (C), smoking (A) is highly correlated with exercise (C): smokers exercise much more than non-smokers. Though smoking causes heart disease, exercise actually is an even stronger preventative measure that one can take against heart disease. The corollary of this is that smoking prevents heart disease (cf. Hitchcock, 2002, p. 9).

The way in which the cause operates may also be unclear. There are many examples one can give here. Morrison (2009, p. 45), for instance, gives the example of small class teaching. In one class operating with small class teaching, the teacher in that class uses highly didactic, formal teaching with marked social distance between the teacher and the student (Factor A), and this is deemed to be an inhibitor of the beneficial effects of small class teaching on students' attainment in mathematics: didactic teaching reduces mathematics performance. However, the same highly didactic, formal class teaching (Factor A) significantly raises the amount of pressure placed on the students to achieve highly (Factor B), and this (Factor B) is known to be the *overriding* cause of any rise in students' performance in mathematics, for example, in small classes the teacher can monitor very closely the work of each child: high pressure raises mathematics performance. Now, it could be argued that Factor A – an ostensibly *inhibiting* factor for the benefits of small class teaching – actually causes *improvements* in mathematics performance in the small class teaching situation.

Another example is where greater examination pressure (A) on students increases their lack of self-confidence (C), but it also increases the student's hard work (B), and hard work reduces the student's lack of self-confidence (C). In other words, the likelihood of the effect of A on C may be lower than the effect of B on C, given A. Let us say that A increases the likelihood of C by 20 per cent, and A increases the likelihood of B by 35 per cent, whilst B reduces the likelihood of C by 75 per cent. In this instance increasing examination pressure increases the student's self-confidence rather than reduces it (see Figure 6.2).

The point here is that a cause might raise the likelihood of an effect, but it may also lower that likelihood, and the presence of other conditions or causes affects the likelihood of an effect of a cause. A diagrammatic representation of these examples is in Figure 6.2 (note that the length of the lines indicates the relative strength of the influence). A cause might lower the likelihood of an effect rather than increase it.

Many cause-and-effect models are premised on linear relationships between cause and effect (i.e. a regular relationship, e.g. a small cause has a regular small effect and a large cause has a regular large effect, or a small cause has a regular large effect and a large cause has a regular small effect). However, seeking linear relations between cause and effect might be misguided, as the effects of causes might be non-linear (e.g. a small or large cause may produce a large, small, irregular or no effect), and it might be to deal with singular or a few causes and singular or a few effects, overlooking the interrelatedness and interactions of multiple causes with each other, with multiple effects and indeed with the multiple interactions of multiple effects. Relationships and their analysis may be probabilistic, conditional and subjunctive rather than linear. Indeed nets and conditions of causation might be more fitting descriptions of causation than causal lines or chains of events or factors (Morrison, 2012).

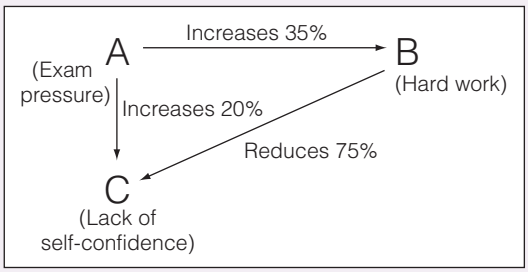


FIGURE 6.2 Positive and negative causes on an effect (1)

One way of focusing on a causal explanation is to examine regularities and then to consider rival explanations of causes and rival hypotheses of these regularities. The observation of regularities, however, is not essential to an understanding of causation, as all cases may be different but no less causative. Further, the best causal explanation is that which is founded on, and draws from, the most comprehensive theory (e.g. that theory which embraces intentionality, agency, interaction as well as structure, i.e. micro- and macro-factors), that explains all the elements of the phenomenon, that fits the *explanandum* (that which is to be explained) and data more fully than rival theories, and which is tested in contexts and with data other than those that have given rise to the theory and causal explanation.

Given the complexity of probabilistic causation, it would be invidious to suppose that a particular intervention will necessarily bring about the intended effect. Any cause or intervention is embedded in a web of other causes, contexts, conditions, circumstances and effects, and these can exert a mediating and altering influence between the cause and its effect.

6.9 The role of action narratives in causation

Statistics, both inferential and descriptive, can indicate powerful relationships. However, these do not necessarily establish unequivocal, direct causation; they may establish the 'what' of causation but not the 'how'. I might assume that A and B cause C, and that C causes D; it is a causal model, and I might measure the effects of A and B on C and the effect of C on D. However, causation here lies in the assumptions behind the model rather than in the statistical tests of the model, and the causal assumptions that lie behind the model derive from theory rather than the model itself (see Chapter 4). Statistics alone do not prove causation. Rather, causation is embodied in the theoretical underpinnings and assumptions that support the model, and the role of statistics is to confirm, challenge, extend and refine these underpinnings and assumptions. Behind statistics that may illuminate causation lie theories and models, and it is in the construct validity of these that causation lies. It is the *mechanisms* of causation – the how and why – that might concern researchers rather than solely numbers and statistical explanations – the what.

Many statistics rely on correlational analysis or on assumptions that pre-exist the statistics, i.e. the statistics might only reinforce existing assumptions and models rather than identify actual causation. Even sophisticated statistics such as structural equation modelling, multiple regression and multivariate analysis succumb to the charge of being no more powerful than the assumptions of causation underpinning them, and, indeed, they often grossly simplify the number or range of causes in a situation, in the pursuit of a simple, clear and easily identifiable model.

How is it that X causes Y; what is happening in X to cause Y? In short, what are the processes of causation? In order to understand this involves regarding causation as dynamic rather than static, as a process rather than a single event, and as involving motives, volitions, reasons, understandings, perceptions, individuality, conditions and context, and the dynamic and emerging interplay of factors, more often than not over time. It is here that qualitative data come into their own, for they 'get inside the head' of the actors in a situation.

A neat example of this is what has come to be known as 'the Rashomon effect' in social sciences (e.g. Roth and Mehta, 2002). It is over sixty-five years since Kurosawa's film *Rashomon* stunned audiences at the Venice Film Festival. It provides four discrepant witness accounts of the same event – an encounter between a samurai, his wife and a bandit, that led to the effect of the samurai's death – in which the causes could have been murder or suicide, consensual sex or rape, fidelity or infidelity. The causal accounts are given by a woodcutter, the bandit, the wife and the spirit of the dead samurai speaking through a medium. Each self-serving account protects the honour of the teller and tries to exonerate each. At the end, there is no clear statement of whose version is correct; truth flounders in the quagmire of epistemology, perception and motives.

Anthropologists, lawyers and social scientists (Roth and Mehta, 2002) seized on the film as an example of the multilayered, contested truth of any situation or its interpretation, coining the term 'the Rashomon effect' to describe an event or truth which is reported or explained in contradictory terms, that gives differing and incompatible causal accounts of an effect: a death. There is more than one causal explanation at work in a situation, and it is the task of the researcher to uncover these, and to examine the causation through the eyes of those imputing the causation.

Action narratives and agency are important in accounting for causation and effects, and, because there is a multiplicity of action narratives and individual motivations in a situation, there are multiple pathways of causation rather than simple input–output models. In understanding the processes of causation, the power of qualitative data is immense, and, indeed, mixed methods may be useful in establishing causation.

Causal explanations that dwell at the level of aggregate variables are incomplete, as behind them, and feeding into them, lie individuals' motives, values, goals and circumstances, and it is these that could be exerting the causal influence; hence a theory of individual motives may be required in understanding and explaining causation.

For example, it is commonplace for a survey to ask respondents to indicate their sex, but it is an entirely different matter – even if different responses are given by males and females to rating scales in a survey – to say that sex *causes* the differences in response. How, actually, is sex a causal factor? Similarly, does social class actually *cause* an effect? It is only a constructed aggregate, a sum of individual characteristics (cf. Kincaid, 2009).

Further, between aggregate independent and dependent variables of cause and effect respectively lie a whole range of causal processes, and these could be influencing the effect and, therefore, have to be taken into account in any causal explanation. How macro-structural features from society actually enter into individuals' actions and interactions, and how individuals' actions and interactions determine social structures – the causal processes involved – need cautious elucidation,

their current status often being opaque processes in a black-box, input–output model of causation.

6.10 Researching causes and effects

The researcher investigating the effects of a cause or the causes of an effect has many questions to answer, for example:

- What is the causal connection between the cause and the effect (how does the cause bring about the effect and how has the effect been brought about by the cause)?
- What are the causal processes at work in the situation being investigated?
- What constitutes the evidence of the causal connection?
- On what basis will the inference of causality be made?
- What constitutes the evidence that a cause is a cause and that an effect is an effect?
- What constitutes the evidence that a cause is the cause (and that there is not another cause) and that an effect is the effect (and that there is not another effect)?
- Is the research investigating the effects of a cause (an interventionist strategy) or the cause of an effect (a post hoc investigation)?
- How will the research separate out a range of possible causes and effects, and how will decisions be made to include and/or exclude possible causes and effects?
- What methodology will be chosen to examine the effects of causes?
- What methodology will be chosen to examine the causes of effects?
- What kind of data will establish probabilistic causation?
- When will the data be collected from which causation will be inferred?

As mentioned earlier, the timing of data collection is a critical feature in establishing causation and the effects of causes. Here the greater the need to establish causal processes, the closer and more frequent should be the data-collection points. Moreover, qualitative data could hold pre-eminence over quantitative methods in establishing causation and causal processes. Further, longitudinal studies might yield accounts of causation that are more robust than cross-sectional studies in which the necessary temporality of causation is built out in favour of the single instance of the data-collection point.

It is not enough to say *that* such-and-such a cause brings about such-and-such an effect, for, whilst it might establish the likelihood *that* the cause brings about an effect or that an effect has been brought about by a cause, this does not tell the researcher *how* the cause brings about the effect or how the effect has been brought about by the cause, i.e. what are the causal processes at work in connecting the cause with the effect and vice versa. If the research really wishes to investigate the processes of causation then this requires detailed, in-depth analysis of the connections between causes and effects.

For example, it is not enough to say that smoking can cause cancer; what is required is to know *how* smoking can cause cancer – what happens between the inhalation of smoke and the presentation of cancer cells. I might say that turning on a light switch causes the light bulb to shine, but this is inaccurate, as turning on the switch completes a circuit of electricity and the electricity causes a filament to heat up such that, when white hot, it emits light.

In education, it is not enough to say that increasing the time spent on reading causes students' reading to improve; that is naive. What might be required is to know how and why the increase in time devoted to reading improves reading. This opens up many possible causes: motivation; concentration levels and spans; interest level of the materials; empathy between the reader and the material; level of difficulty of the text; purposes of the reading (e.g. for pleasure, for information, for learning, for a test); reading abilities and skills in the reader; subject matter of the text; ambient noise; where, when and for how long the reading is done; prior discussion of, and preparation for, the reading material; follow-up to the reading; choice of reading materials; whether the reading is done individually or in groups; teacher help and support in the reading time; relatedness of the reading to other activities; the nature, contents and timing of the pre-test and post-test; the evidence of improvement (and improvement in which aspects of reading); and so on.

It can be seen in this example of reading that the simple input variable – increasing time for reading – may bring about an improvement in reading, but that may only be one of several causes of the improvement, or an umbrella term, or may liberate a range of other causes to come into play, both direct and indirect causes. Identifying the true cause(s) of an effect is extremely difficult to pin down.

Take, for example, the introduction of total quality management into schools. Here several interventions are introduced into a school for school improvement, and, at the next school inspection, the school is found

to have improved. The problem is trying to decide which intervention(s) has/have brought about the improvement, or which combinations of interventions have worked, or which interventions were counter-productive, and so on. It is akin to one going to the doctor about a digestion problem; the doctor prescribes six medicines and the digestion problem goes. Which medicine(s) was/were responsible for the cure, and in what combinations, or is it really the medicines that have brought about the cure; were there other factors that brought about the cure; would the digestion problem have cured itself naturally over time?

The researcher has to identify which cause (A) or combination of causes have brought about which effect (B), both intended and unintended, or whether the supposed cause (A) brought about another effect (C) which, in turn, became the cause of the effect (B) in question, and whether the effect (B) is really the consequence of the supposed cause(s) (A), and not the consequence of something else. What looks like being a simple cause-and-effect actually explodes into a multiplicity of causes and effects (Figure 6.3) (cf. Morrison, 2009, p. 124).

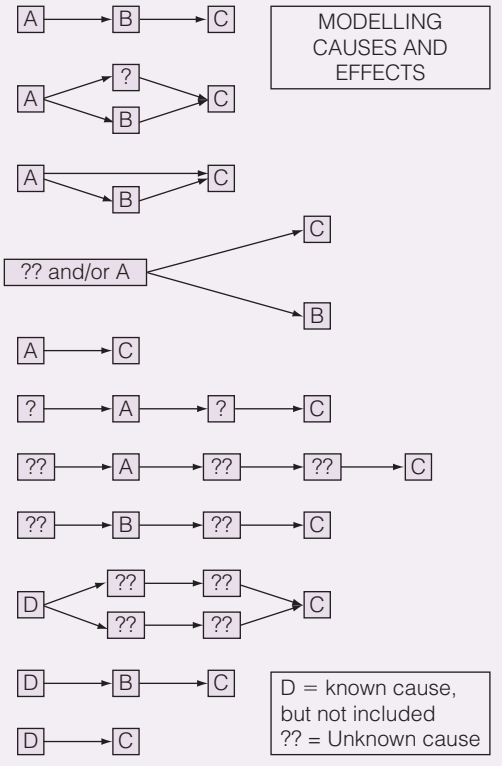


FIGURE 6.3 Positive and negative causes on an effect (2)

How, then, can the researcher proceed in trying to uncover causes and effects? A main principle underpinning how some researchers operate here is through *control*, isolating and controlling all the variables deemed to be at work in the situation. By such isolation and control, one can then manipulate one or more variables and see the difference that they make to the effect. If all the variables in a situation are controlled, and one of these is manipulated, and that changes the effect, then the researcher concludes that the effect is caused by the variable that has been manipulated. Moreover, if the research (e.g. an experiment) can be repeated, or if further data (e.g. survey data) are added, and the same findings are discovered, then this might give added weight to the inferred cause-and-effect connection (though regularity – Hume's (2000) 'constant conjunction' – is no requirement for causation to be demonstrated). This assumes that one has identified, isolated and controlled all the relevant variables, but, as the earlier part of this chapter has suggested, this may be impossible.

One way in which the problem of isolation and control of variables is addressed is through randomization – a key feature of the 'true' experiment (see Chapter 20). For example, random allocation of individuals to a control group or an experimental group is a widely used means of allowing for the many uncontrolled variables that are part of the make-up of the groups in question (Schneider *et al*., 2007). It adopts the *ceteris paribus* condition (all other things being equal) that assumes that these many other variables are evenly distributed across the groups, such that there is no need to control for them. This is a bold and perhaps dangerous assumption to make, not least as chaos and complexity theory tell us that small changes and differences can bring about major differences in outcome.

Whilst control is one prime means of trying to establish causation, it does raise several problems of the possibility, acceptability or manageability of isolating and controlling variables, of disturbing and distorting the real work of the participants, and of operating an undesirable – even unethical – control and manipulation of people. This is the world in which the researcher is king or queen and the participants are subjects – subjected to control and manipulation. On the one hand the claim is made that the research is 'objective', 'clean' (i.e. not affected by the particular factors within each participant), laboratory-based and not prone to bias; on the other hand it is a manipulative and perhaps unrealistic attempt to control a world that cannot in truth be controlled. Are there alternatives?

A major alternative is one that keeps the 'real' world of participants as undisturbed as possible, avoids the

researcher controlling the situation, and uses qualitative data to investigate causation. Here observational, interview and ethnographic methods come to the fore, and these are very powerful in addressing the *processes* of causation and in establishing the causes of an effect as recounted by the participants or the observers themselves. These methods deliberately 'get inside the heads' of individuals and groups, as well as including the researcher's own views, identifying and reporting causation in their terms. They provide considerable authenticity to the causal accounts given or compile a sufficiently detailed account of a situation for the researcher to make informed comments on the workings of causation in the situation under investigation. Further, it is often the participants themselves who identify what are the causes of effects in the situations being investigated (though the researcher would need to be assured that these are genuine, as participants may have reasons for not disclosing the real causes or motives in a situation or, indeed, may be mistaken).

These two approaches are not mutually exclusive in a piece of research, and, as Chapter 2 has indicated, there is an advantage in adopting a mixed methods approach, or, indeed, in a mixed methodology approach, in which positivist and experimental approaches might yield accounts of the 'what' of causation – which variables are operating to produce an effect – whilst an interpretive approach might be used to yield data on the 'how' and 'why' of causation – how the causal processes are actually working.

Researchers examining causes and effects have to decide whether they are researching the effects of causes (e.g. in which they introduce an intervention and see what happens as a consequence) or the causes of effects (e.g. backtracking from an observed situation to try to discover its causes). These are discussed below.

6.11 Researching the effects of causes

In trying to investigate the effects of one or more causes, the researcher can commence with a theory of causality operating in a situation (e.g. bringing pressure to bear on students causes them to work harder, or dropping out of school reduces income at age 50 by a factor of five, or improving self-esteem improves creativity), operationalize it, and then test it, eliminate rival theories and explanations using data other than those which gave rise to the explanation, and then proceed to the drawing and delimiting of conclusions. The use of continuous rather than categorical variables might be more effective in establishing the nature and extent of causation as they indicate the magnitude of causes and effects.

On the other hand, the researcher can proceed along an entirely different track, using qualitative research to really understand the causal processes at work in a situation and in the minds of the participants in that situation – the 'how' and 'why' of causation.

Determining the effects of causes is often undertaken using an interventionist strategy in educational research, installing an intervention either to test a hypothesized causal influence or a causal model, or because it is already known that it may exert a causal influence on effects, i.e. manipulating variables in order to produce effects. (Of course, a non-intervention may also be a cause, for example, I may cause a plant to die by not watering it, i.e. by doing nothing.)

Manipulation takes many forms, including:

- action research (discussed in Chapter 22), but this may raise questions of rigour brought about by a lack of controls and a lack of external checks such that the attribution of causation may be misplaced;
- a range of experimental approaches (discussed in Chapter 20), which assume, perhaps correctly or incorrectly, acceptably or unacceptably, that variables and people can be isolated, controlled and manipulated; and
- participant observation in qualitative research.

In addressing these approaches, however, serious attention has to be paid to a range of issues:

- the context of the intervention and the power of the situation could affect the outcomes and behaviours of participants (the Hawthorne effect or the Lucifer effect (Zimbardo, 2007a));
- the same causes do not always produce the same effects, even with the same people;
- inappropriate timing of the pre-test and post-test measurements of effects could undermine the reliability of the statement of the effects of the cause;
- there are problems of accuracy and reliability, as groups and individuals cannot both be in a group that is and is not receiving an intervention (Holland's (1986, p. 947) 'fundamental problem of causal inference', which may not be sufficiently attenuated by randomization) (see Chapter 20);
- process variables and factors, and not only input variables, as these feature in understanding causation;
- the characteristics, personae and specific individual features of participants and their agency, as these influence interventions and their effects.

Experimental techniques, particularly randomized controlled trials (RCT), have some potency in establishing

causation, and it is here that the identification, isolation and control of independent variables is undertaken, manipulating one independent variable to see if it makes a difference to the outcome. The other variables are held constant and, if a change of outcome is found by manipulating the one independent variable, then the change can be attributed to that independent variable (it becomes the cause), as the other variables have been held constant, i.e. their influence has been ruled out.

In experimental approaches, randomization is an important element in determining causation in order to overcome the myriad range of variables present in, and operating in, participants (the *ceteris paribus* condition discussed earlier), to overcome within-group and between-group differences (cf. Fisher's *The Design of Experiments* (1966)). RCTs and experiments (see Chapter 20) are an example of interventionist approaches that seek to establish the effects of causes by introducing one or more interventions into a situation and observing the outcomes of these under controlled conditions.

However, RCTs are often not possible in education and, indeed, are not immune to criticism. For example, the assumptions on which they are founded may be suspect (e.g. over-simplifying the variables at work in a situation, and overriding the influence of mediating or process variables). They may have limited generalizability (Cartwright and Hardie, 2012) and the measures used in RCTs focus on average results rather than outliers or important sub-sample differences. They frequently do not establish the causal processes or causal chains that obtain in the situation. They neglect participants' motives and motivations. They neglect the context in which the action is located, and they might neglect the moral agency of participants and the ethics of researchers. Indeed context can exert a more powerful causal force than the initial causal intervention, as evidenced in the examples of the Stanford Prison Experiment and the Milgram experiments on obedience (see Chapters 7 and 30).

Caution must be exercised in supposing that RCTs, for some people the epitome of causal manipulation in the determination of the effects of causes, will yield sufficient evidence of causation, as these overlook the significance of context and conditions, of processes, of human intentionality, motives and agency, over-determination etc., in short, of the contiguous causal connections between the intervention and its putative effects. Indeed even the issue of when and whether an effect has an effect (short-term to long-term, immediate or delayed) is problematic, and attention had to be given to effects that have been caused by the intervention other than those in which the researcher might be initially interested. For example, a researcher might find that pressuring students to learn improves their mathematics scores but leads to an enduring dislike of mathematics.

In the context of moves towards judging 'what works', deciding 'what works' is as much a matter of values and judgement as it is of empirical outcomes of causation. Success is a value judgement, not simply a measure or a matter of performance. Judging 'what works' in terms of cause and effect is an incomplete analysis of the situation under investigation. A more fitting question should be 'what works for whom, under what conditions, according to what criteria, with what ethical justifiability, and with what consequences for participants?'

A range of issues in judging the reliability and validity of experimental approaches in establishing causation includes the acceptability of laboratory experiments that are divorced from the 'real world' of multiple human behaviours and actions. Here field experiments and natural experiments (see Chapter 20) may attenuate the difficulties posed by laboratory experiments, though these, too, may also create their own problems of reliability and validity.

As an alternative to action research and experimental methods in determining effects from causes, observational approaches can be used, employing both participant and non-participant approaches (see Chapter 26). Whilst these can catch human intentionality, agency and perceptions of causality and events more fully than experimental methods, nevertheless they encounter the same difficulty as action research and experiments, as they, too, have to provide accounts of causal processes and causal chains. Further, in addressing intentionality and agency in causal processes and chains, it is also possible that, whilst perceptions might be correct, they might also be fallacious, partial, incomplete, selective, blind and misinformed. I might think that there is a mouse in the room (a cause), and act on the basis of this (an effect), but, in fact, there may be no mouse at all in the room.

Interventionist approaches, and the determination of the effects of causes, risk mixing perception with fact, and, regardless of evidence, human inclinations may be to judge data and situations on the basis of personal perceptions and opinions that, indeed, may fly in the face of evidence (the 'base rate fallacy'; Morrison, 2009, pp. 170–1; see also Kahneman, 2012). This is only one source of unreliability, and it is important to consider carefully what actually are the effects of causes rather than jumping to statements of causation based on premature evidence of connections.

6.12 Researching the causes of effects

Determining the causes of effects is even more provisional, tentative and inferential than determining the effects of causes, as data are incomplete and backtracking along causal chains and/or searching within causal nets is difficult, as it requires a search for clues and testing rival hypotheses about causation. It is possible to generate a huge number of potential causes of observed effects, and the problem is in deciding which one(s) is/are correct. Morrison (2009) suggests that one approach which can be adopted in tracing causes from effects is *ex post facto* research (see Chapter 20), but it poses challenges in the sometime inability to control and manipulate independent variables or to establish randomization in the sample. Another approach is to adopt a seven-stage process of tracing causes from effect, thus:

Stage 1: Establish exactly what has to be explained.
Stage 2: Set out possible theoretical foundations for the investigation.
Stage 3: Examine, evaluate and eliminate rival theoretical foundations, selecting the most fitting.
Stage 4: Hypothesize a causal explanation on the basis of the best theoretical foundation.
Stage 5: Set out the assumptions underlying the causal explanation.
Stage 6: Test the causal hypotheses empirically.
Stage 7: Draw conclusions based on the test.

A worked example is provided here, from Goldthorpe (2007). Goldthorpe seeks to explain the causes of 'persistent differentials in educational attainment' despite increased educational expansion, provision and uptake across the class structure (p. 21), i.e. in the context of increased educational opportunity and its putative weakening influence on class-based determination of life chances. He proceeds in the seven stages indicated above. Only after that test does he provide a causal explanation for his observed effects.

Stage 1: Establish what it is that has to be explained

First, Goldthorpe observes some 'regularities' (effects) (2007, p. 45):

a In all economically advanced societies there has been an expansion over time of education provision and in the numbers of students staying on in full-time education beyond the minimum schooling age (e.g. going into higher education).

b At the same time, class differentials in educational attainment have remained stubbornly stable and resistant to change, i.e. though students from all classes have participated in expanded education, class origins and their relationship to the likelihood of them staying on in education or entering higher education has only reduced slightly, if at all, and this applies to most societies.

He is establishing social regularities that any causal and theoretical account should seek to explain: the creation, persistence and continued existence of class stratification in modern societies, and the continuing class-relatedness of educational inequality and life chances (p. 24).

Stage 2: Set out possible theoretical foundations for the investigation

Goldthorpe's work is premised on the view that theories are necessary to provide explanatory foundations for how established regularities come to be as they are (2007, p. 21). He initially suggests four theoretical foundations: Marxist theory, liberal theory, cultural theory and rational choice theory.

Stage 3: Examine, evaluate and eliminate rival theoretical foundations

For several reasons which he gives (2007, pp. 22–34), Goldthorpe rejects the first three of these and argues that rational choice theory provides a fitting theoretical foundation for his investigation of the causes of the effects observed (pp. 34–41). True to rational action theory, Goldthorpe places emphasis on aspirations, in particular noting their relative rather than their absolute status, that is to say, aspirations are relative to class position, as working-class aspirations may not be the same as those of other classes (p. 31). Different social classes have different levels and kinds of aspiration, influenced – as rational action theory suggests – by the constraints under which they operate, and the perceived costs and benefits that obtain when making decisions (p. 32). Taking *relative* rather than absolute views of aspiration enables accounts to be given that include the fact of increased provision of, and participation in, education by students from all social classes, i.e. class differentials have not widened as education provision and participation have widened.

Goldthorpe (2007, p. 32) suggests that cultural theory may account for what Boudon (1973) terms 'primary effects', i.e. initial levels of achievement and ability in the early stages of schooling. However, he is more concerned with Boudon's 'secondary effects',

i.e. those effects which come into play when children reach 'branching points' (transition points, e.g. from primary to secondary schooling, from secondary education to university) (p. 32) and which have increasingly powerful effects as one progresses through schooling. 'Secondary effects' take account of the aspirations and values that children and their parents hold for education, success and life options, i.e. the intentionality and agency of rational action theory in a way that 'primary effects' do not. Goldthorpe notes that, at each successive 'branching point' (p. 32), children from more advantaged backgrounds remain in the educational system and those from less advantaged backgrounds either leave school or choose courses that lead to lower qualifications (hence reducing their opportunities for yet further education).

Goldthorpe (2007, p. 33) argues that more ambitious options may be regarded less favourably by those from less advantaged class backgrounds as they involve: (a) greater risk of failure; (b) greater cost; and (c) relatively less benefit. In other words, the level of aspiration may vary according to class and the associated levels of assessed cost and risk by members of different classes, and children from less advantaged backgrounds have to be more ambitious than those from more advantaged backgrounds if they are to meet the aspirations and success levels of those from more advantaged backgrounds. Class origins influence risk assessment, cost assessment and benefit assessment – all aspects that are embraced in rational action theory. These determine the choices made by children and their parents.

Stage 4: Hypothesize a causal explanation on the basis of the best theoretical foundation

Goldthorpe (2007, p. 34) argues that class differentials in educational attainment have persisted because, even though there has been expansion and reform of education, and even though the overall costs and benefits that are associated with having more ambitious options have encouraged their take-up, in practice there has been little concurrent change in the 'relativities between *class-specific* balances': different classes view the costs, risks and benefits differently (p. 34). This is his working hypothesis in trying to establish cause from effect.

Stage 5: Set out the assumptions underlying the causal explanation

Goldthorpe tests his theory by drawing initial attention to the ongoing income differentials between classes; indeed he argues that they have widened (2007, p. 35), with manual labourers more prone to unemployment than professional or managerial workers, i.e. the costs of education are still a factor for less advantaged families, particularly at the end of the period of compulsory schooling. At the time when their children come to the end of compulsory schooling, the income of manual workers will already have peaked (e.g. when they are in their forties), whereas for professional and managerial workers it will still be rising, i.e. costs are more of a problem for manual workers than for professional and managerial workers, i.e. the costs of higher education relative to income, and the consequent effects on family lifestyle if families are having to finance higher education, are much higher for manual workers. This increased proportion of family income to be spent on education for less advantaged families is coupled with the fact that, if children from these families are to succeed, then they need even more ambition than their professional and managerial class counterparts, i.e. they are at a potential double disadvantage, i.e. relative advantage and disadvantage are not disturbed, a feature on which liberal theory is silent (p. 36).

Goldthorpe makes the point that class position conditions educational decisions made by members of different classes. These different class positions influence different evaluations of the costs and benefits of education, and these are socially reproductive, i.e. the social class position is undisturbed.

Another element of his argument concerns risk aversion. His view is that a major concern of members of different classes is to minimize their risk of downward class mobility, and to maximize their chances for upward class mobility or, at least, to maintain their existing class location (p. 37). This exerts greater pressure on the already-advantaged classes (e.g. the salariat) to have their children complete higher education (in order to preserve intergenerational class stability) than it does on the children from less advantaged classes (e.g. the waged). It costs more for the children of the advantaged classes to preserve their class position than it does for children of the less advantaged classes to preserve theirs.

With regard to families in the less advantaged classes, Goldthorpe (p. 38) suggests that they regard higher education much more guardedly. Not only does it cost less for them to maintain their class position, but it costs relatively more to achieve upward class mobility; their best options might be for vocational education, as it is cheaper and gives a strong guarantee of *not* moving downwards in class situation (e.g. to be unemployed or unskilled).

Further, for children in this class, the costs (and likelihood) of failure in higher education could be

proportionately greater than those for children from more advantaged families. For example, in terms of: the relative costs of the higher education; lost earning time; lost opportunity to follow a vocational route in which they have greater likelihood of being successful (p. 38); loss of social solidarity if working-class children pursue higher education, the consequences of which may be to remove them from their class origin and community (pp. 38–9). These factors combine to suggest that children and families from less advantaged backgrounds will require a greater assurance, or expectation, of success in higher education before committing themselves to it than is the case for children and families from more advantaged backgrounds (p. 68).

Goldthorpe then offers his causal explanation of the effects observed: the persistence of class differentials in educational attainment despite expansion of educational provision and participation (p. 39):

1. Class differentials in the uptake of more ambitious educational options remain because the conditions also remain in which the perceived costs and benefits of these options operate, and these lead to children from less advantaged families generally requiring a greater assurance of success than children from more advantaged families before they (the former) pursue more ambitious educational options;
2. There is a rational explanation for the persistence of these different considerations of ambitious options by class over time, which is rooted in class-based conditions.

These are the two main hypotheses that he seeks to test.

Stage 6: Test the hypotheses empirically

Goldthorpe (2007) then proceeds to test his two hypotheses (pp. 39–44, 53–6, and his chapters 3 and 4), adducing evidence concerning several factors, for example:

- the greater sensitivity of working-class families to the chances of success and failure in comparison to middle-class families (p. 40);
- different levels of ambition in working-class and middle-class families (p. 40);
- relative (class-based) risk aversion in decision making: for example, the risk of failure and/or of closing options (pp. 55–6);
- the loss of forgone earnings (pp. 53–5);
- expectations of success (pp. 55–6);
- evaluation of the potential benefits, value and utility of higher education (pp. 38–9);
- influences on choices and decision making in different classes (his chapter 3);
- actual choices made by members of different classes;
- fear of downward social mobility (pp. 53–4);
- the need to preserve, or improve on, intergenerational mobility (pp. 53–4);
- financial costs (p. 56).

He indicates that students from lower socio-economic groups either cannot afford, or cannot afford to take risks in, higher education, and he identifies three clusters of possible explanations of persistence of class differentials in educational attainment, including (but not limited to):

Cluster 1: Differences in aspirations and decisions are caused by perceptions of costs: (a) loss of earnings during study time (a bigger drawback for families and students from low-income households than for those from privileged backgrounds); (b) students from low-income households have to work harder than privileged students in order to compete with them; (c) students from low-income households must have greater ambition than privileged students in order to be successful in a higher social class; (d) the financial costs of higher education, proportional to income, are higher for less advantaged students than for more advantaged students and families.

Cluster 2: Differences in aspiration and decisions are caused by relative risk aversion: (a) the risk of failure in higher education is greater for students from disadvantaged classes; (b) the risk of loss of further educational opportunities if failure ensues or incorrect options are followed is greater for students from disadvantaged classes than for students from more privileged classes; (c) the risk of loss of social solidarity is greater for students from working-class groups than for students from more privileged classes; (d) less advantaged students must have greater ambition than privileged students in order to be successful in higher social classes.

Cluster 3: Differences in aspiration and decisions are caused by perceptions of relative benefit: (a) the opportunity for upward social mobility through higher education is an attraction for students from lower-class backgrounds; (b) higher education is differentially necessary for preferred or likely employment for those from privileged and less privileged groups.

Stage 7: Draw conclusions based on the test

Goldthorpe (2007) indicates that class differentials have continued to affect the take-up of educational options. He finds that class differentials in terms of the take-up of more ambitious educational options have

been maintained because so too have the conditions in which the perceived costs and benefits of these options lead to children from less advantaged families requiring, on average, a greater assurance of success than their more advantaged counterparts before they decide to pursue such options. There are class differences in terms of relative ambition, risk aversion, perceived costs and benefits, amounts of effort required, assurances of success (and the significance of this), fear of downward social mobility, income, occupational choices and the need for qualifications.

He concludes that the results of empirical tests support his explanation of the factors of relative risk aversion and fear of downward social mobility exerting causal power on educational decision making which, in turn, lead to class differentials in educational attainment being maintained (p. 99).

Goldthorpe argues that this hypothesis is better supported than alternative hypotheses (e.g. educational choices being predetermined by culture, class identity and the class structure).

This lengthy example here offers a robust account of how to track backwards from an effect to a cause and how to evaluate the likelihood that the putative cause of the effect actually is the cause of that effect. In summary, for researchers seeking to establish the causes of effects, the task has several aspects:

- Indicate what needs to be done to test the theory and to falsify it.
- Identify the kinds of data required for the theory to be tested.
- Identify the actual data required to test the theory.
- Identify the test conditions and criteria.
- Construct the empirical test.
- Consider the use of primary and secondary data.
- Consider using existing published evidence as part of the empirical test.
- Ensure that action narratives and intentionality are included in causal accounts.

The fundamental problem in determining causes from effects is the uncertainty that surrounds the status of the putative cause; it can only ever be the best to date, and the researcher does not know if it is the best in absolute terms. One effect stems from many causes, and to try to unravel and support hypotheses about these may present immense difficulties for the researcher. Morrison (2009, p. 204) suggests that there are several ways in which causes may be inferred from effects:

- recognizing that a high level of detail may be required in order to establish causation: high granularity;
- identifying several causal chains, mechanisms and processes in a situation;
- combining micro- and macro-levels of analysis;
- addressing both agency and structure;
- underpinning the data analysis and causal explanation with theory;
- using different kinds of ex post facto analysis;
- using correlational and causal-comparative, criterion group analysis;
- ensuring matching of groups in samples and that similar causes apply to both groups;
- adopting the seven-stage process set out above, of generation, testing and elimination of hypotheses and rival hypotheses;
- ensuring clarity on the direction of causation;
- using empirical data to test the causal explanation;
- identifying which is cause and which is effect, and/or which effect then, subsequently, becomes a cause;
- avoiding the problem of over-selective data;
- ensuring that the data fairly represent the phenomenon under investigation;
- recognizing that cause and effect may be blurred;
- accepting that effects may become causes in a cyclical sequence of causation;
- seeking out and recognizing over-determination at work in causal accounts;
- keeping separate the explanans (the explanation) from the explanandum (that which is to be explained);
- ensuring that alternative theories and causal explanations are explored and tested;
- drawing conclusions based on the evidence, and the evidence alone.

In seeking to establish the causes of effects, there is a need to review and test rival causal theories and to retain those with the greatest explanatory potential and which fit the evidence most comprehensively and securely. Testing rival hypotheses must be done with data that are different from those that gave rise to the hypotheses, in order to avoid circularity.

The determination of causes from effects does not have the luxury afforded to causal manipulation in determining effects of causes. Whilst this renders the determination of causes from effects more intractable, nevertheless this is not to say that it cannot be attempted or achieved, only that it is difficult.

Morrison (2009) argues that, in seeking to identify the causes of effects, there is a need for a theoretical foundation to inform causal explanation. Possible causal explanations should be evaluated against rival theories and rival explanations, being operationalized

in considerable detail (high granularity), and tested against data that are different from those that gave rise to the causal explanation. Causal explanations should link micro- and macro-factors, include agency and intentionality as well as structural constraints, and contain a level of detail that is sufficiently high in granularity to explain the phenomenon to be explained without concealing or swamping the main points with detail overload, i.e. the researcher must be able to distinguish the wood from the trees.

The companion website to the book presents a fully worked example of working with a range of challenges in causality, for example, counterfactuals, 'before-and-after' comparisons, multiple causes, over-determination, causal forks, preceding causes, causal links, causal direction and what can and cannot be inferred about causality. We advise readers to look at that in-depth worked example.

6.13 Conclusion

In approaching causal research, then, the researcher is faced with a range of challenges, including, for example:

- focusing more on causal processes than input/output/results models of causation;
- establishing causation other than through reduction and recombination of atomistic, individual items and elements;
- regarding causation as the understanding of the emergent history of a phenomenon or a whole;
- investigating multiple and simultaneous causes and their multiple and simultaneous effects in a multiply connected and networked world;
- separating causation from predictability, and drawing the boundaries of predictability for an understanding of the frequent uniqueness of a causal sequence, which may not be repeatable, i.e. living with uncertainty and unpredictability;
- learning to work with causation in a situation in which randomness often 'trumps' causation (cf. Gorard, 2001a, p. 21);
- indicating the utility of understanding causation if it has little subsequent predictive strength;
- understanding how to investigate causation in holistic webs of connections, i.e. how is it possible to discover or demonstrate causation when looking at events holistically;
- understanding causation and causal processes in a multi-causal, multi-effect, non-linear and multiply connected world;
- identifying the causal processes at work in determining social and macro-structures from the actions and interaction of individuals (the micro-worlds) and, conversely, in determining the actions and interactions of individuals from the structures of society and its institutions (the macro-worlds), their ontologies and epistemologies.

The researcher has to decide whether the research is investigating the cause of an effect, the effect of a cause, or both, and when causation is demonstrated, given that absolute certainty is illusory. If one is investigating the effects of causes then the methodologies and approaches to be used might include experiments, action research, survey analysis, observational approaches or a combination of these (and indeed others). If one is investigating the causes of effects then, in the context of the likelihood of greater uncertainty than in establishing the causes of effects, one can employ numerical and qualitative data in backtracking from effects to causes and in testing hypothesized causes of effects. In all of these approaches, this chapter has suggested that probabilistic rather than deterministic causation is a more fitting description of the nature of the conclusions reached. It has suggested that, even if it sounds simplistic at first, nevertheless it is both important yet difficult to establish what actually constitutes a cause and an effect. The chapter has suggested that causal processes, with high granularity, are often closer to identifying the operations of causes and effects and the links between them, and that here qualitative data might hold pre-eminence in educational research. However, the chapter has also suggested that there is an important role for numerical approaches, for examining the 'regularities' that might be evidenced in survey approaches, and in the isolation and control of variables in experimental approaches. In short, the chapter is arguing for the power of mixed methodologies and mixed methods in investigating and establishing causation.

None of the preceding discussion takes us very far from the difficulty in actually defining causation. Is it, like time, space, existence, not defined in terms of previously defined concepts, hence is a 'primitive concept', irreducible to anything else?

The companion website to the book provides additional material and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: www.routledge.com/cw/cohen.

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Part 2
Research design

The planning of educational research is not an arbitrary matter; the research itself is an inescapably ethical enterprise. Nor are the planning and conduct of educational research simply a matter of cranking out recipes and following them. On the contrary, educational research is a deliberative and reflexive exercise. We place ethical issues at a very early point in the book to signal this. The research community and those using the findings have a right to expect that research is conducted rigorously, scrupulously and in an ethically defensible manner. This is thrown into sharp relief with the rise in online research, and we discuss this in a new chapter. All this necessitates careful planning, and this part introduces some key design and planning issues. It contains chapters on how to choose a research project and a comprehensive set of considerations in the design and planning of educational research, including ensuring that the research provides warrants for interpreting data and drawing conclusions. This part includes another entirely new chapter on research questions and hypotheses. The entire part contains a wide range of worked examples.

In designing research, we need to consider the issues of how to choose a research project, how to plan it, how to conduct a literature search and review, and how to ensure that the project is practicable. This part suggests several ways that researchers can approach the choice of a research project, and comments on the need for the project to be significant (and what this means), to consider its purposes and intended outcomes, feasibility, research questions, literature review and overall design.

This part also provides an augmented chapter on sampling issues, with attention to statistical power. Sampling, reliability and validity are key matters in research; without due attention to these the research could turn out to be worthless. Hence this part addresses these issues in detail. These are complex matters, and we take readers through them systematically. The chapter on sensitive educational research is included here, to underline the point that not only is the very decision to conduct research a sensitive matter, but that often access itself is difficult and sensitive, and this could be the major issue to be faced in planning research. This part sets out a range of planning possibilities so that the eventual selection of sampling procedures, together with decisions on reliability and validity, are made on the basis of *fitness for purpose*, and so that sensitivities in research are anticipated and addressed.

The ethics of educational and social research

CHAPTER 7

7.1 Introduction

Ethics concerns that which is good and bad, right and wrong. Ethical research concerns what researchers ought and ought not to do in their research and research behaviour. A cursory glance at recent literature throws up a vast field of issues in considering ethics in educational research, for example:

- informed consent;
- confidentiality and anonymity;
- identification and non-traceability;
- non-maleficence;
- beneficence and duty of care;
- responsibilities (for what and to whom);
- gaining access;
- overt and covert research;
- disclosure and public versus private knowledge and spaces;
- relationships and differential power relations in research;
- interests at stake in the research (in whose interests the research is operating);
- rights, permissions and protections;
- ownership and control of data;
- access to data (and its archiving);
- the roles and power of research sponsors and commissioners;
- sensitive research;
- gender, age, colour, (dis)ability and ethnicity issues;
- researching with children;
- avoidance of selective, partisan and skewed data analysis;
- value positions in data interpretation;
- responsibilities to different parties;
- being judgemental.

Each of these, in turn, raises many questions and considerations and we introduce them in this chapter. There are rarely easy, 'black-and-white' decisions on ethical matters. Rather, researchers must take informed decisions on a case-by-case basis. Codes of practice, ethical guidelines, ethics committees and institutional review boards, legislation, regulations and regulatory frameworks may raise issues for consideration and provide advice for researchers on what to do and not to do. However, ethical issues are rarely as straightforward as rule-following would suggest, and it is for individuals to take responsibility for the decisions that they take on ethical matters and the actions connected with those decisions (Brooks *et al*., 2014, p. 153).

Ethical decisions are contextually situated – socially, politically, institutionally, culturally, personally – and each piece of research raises ethical issues and dilemmas for the researcher. Ethical norms vary in different parts of the world, and what is acceptable in a western culture may not apply elsewhere. Ethical issues are not a once-and-for-all matter which can be decided before the research commences or when the proposal is put to an ethics committee, and then forgotten (cf. Brooks *et al*., 2014, p. 154); rather, they run throughout the entire research process. For example, Wax (1982, p. 42) makes the telling point that informed consent in many kinds of research is not a 'one-shot, once-and-for-all' affair, but has to be continuously negotiated, particularly in qualitative, emergent research. Ethics are present at every turn, and we indicate key issues to be faced at each stage.

What starts as being an apparently straightforward ethical matter quickly raises non-straightforward ethical decisions for the researcher. Each research undertaking is an event *sui generis*, and the conduct of researchers cannot be, indeed should not be, forced into a procrustean system of ethics. When it comes to the resolution of a specific moral problem, each situation frequently offers a spectrum of possibilities. Ethics are 'situated', i.e. they have to be interpreted in specific, local situations (Simons and Usher, 2000).

Each stage in the research sequence raises ethical issues. Sikes (2006) notes that ethics touch 'researchers and their research choices, research topics, methodologies and methods, and writing styles' (p. 106). Ethical issues may arise from the nature of the research project itself; the context for the research; the procedures to be adopted (e.g. creating anxiety); methods of data collection (e.g. covert observation); the nature of the

participants; the type of data collected (e.g. personal and sensitive information); what is to be done with the data (e.g. publishing in a manner that may cause participants embarrassment or harm); and reporting the data (e.g. in a way that the participants will understand) (Oliver, 2003, p. 17).

How, then, can the researcher, particularly the novice researcher, begin to address the scope of ethical issues? One way is to follow the stages of research, from initial considerations to research planning, choice of topic, design, methodologies, data collection, data analysis, interpretation, to reporting and dissemination, and this is how the chapter is organized. We review issues in the ethical field in the sequence in which they may be encountered in planning, conducting, reporting and disseminating research:

- ethical principles and the nature of ethics in educational research;
- sponsored research;
- regulatory contexts of ethics: codes of practice, ethical review boards and ethics committees, legislation, ethical frameworks and guidelines;
- choice of research topic and research design;
- ethical dilemmas in planning research: informed consent, non-maleficence, beneficence and human dignity, privacy, anonymity, confidentiality, betrayal and deception;
- gaining access and acceptance into the research setting;
- power and position;
- reciprocity;
- ethics in data analysis;
- ethics in reporting and dissemination;
- responsibilities to sponsors, authors and the research community.

These are intended to guide the reader through a maze of ethical concerns in educational research, and the foundations on which they are built. The chapter provides practical examples of ethics considerations.

7.2 Ethical principles and the nature of ethics in educational research

Ethics has been defined as 'a matter of principled sensitivity to the rights of others' (Cavan, 1977, p. 810). Educational researchers must take into account the effects of the research on participants; they have a responsibility to participants to act in such a way as to preserve their dignity as human beings. Ethical decisions are built on ethical principles, but different ethical principles may conflict with each other, and we explore this below.

Ethical problems in educational research can often result from thoughtlessness, oversight or taking matters for granted. A student whose research is part of a course requirement and who is motivated wholly by self-interest, or academic researchers with professional advancement in mind, may overlook the 'oughts' and 'ought nots'. It is unethical for the researcher to be incompetent in the area of research. Competence may require training (Ticehurst and Veal, 2000, p. 55). Indeed an ethical piece of research must demonstrate rigour and quality in the design, conduct, analysis and reporting of the research (Morrison, 1996b).

Kimmel (1988) has pointed out that it is important we recognize that the distinction between ethical and unethical behaviour is not dichotomous, even though the normative code of prescribed ('ought') and proscribed ('ought not') behaviours, as represented by the ethical standards of a profession, seem to imply that it is. Judgements about ethics lie on a *continuum* that ranges from the clearly ethical to the clearly unethical. The point here is that ethical principles are not absolute, generally speaking (though some may maintain otherwise), but must be interpreted in the light of the research context and of other values at stake.

Whilst many of the issues addressed in this chapter concern procedural ethics, ethics concerns right and wrong, good and bad, and so procedural ethics may not be enough; one has to consider how the research purposes, design, contents, methods, reporting and outcomes abide by ethical principles and practices.

A *deontological* view of ethics concerns what one has a duty or obligation to undertake, what ought to be done, i.e. there are universalizable 'categorical imperatives' (Kant's phrase) to behave in certain ways and not to behave in other ways (prescriptive and prohibitive respectively), which override the consequences of such actions. These are universalizable in the sense that it should apply to all persons, and categorical in that they must be obeyed without exception. This view involves treating people as ends in themselves – with equal respect, dignity and value – rather than as means (Howe and Moses, 1999, p. 22). As Brooks *et al*. (2014, p. 23) remark, the deontological view requires us to treat others as we would wish others to treat us, regardless of their personal characteristics, status or backgrounds. This extends to considerations such as: do no harm and do prevent harm (non-maleficence); do good (beneficence); be honest; be sincere; be grateful. We discuss these issues in detail later in the chapter.

By contrast, a *consequentialist* view of ethics concerns the outcomes of actions, for example, the utilitarian view that ethical behaviour is that which produces the greatest good (and happiness) for the greatest

number. This replaces the deontological emphasis on unexceptionable rules and what is always right with a focus on consequences. In this view a costs–benefits analysis is considered, but this is problematical, as (a) it is unclear which costs and which benefits should be factored into the analysis, and it assumes that all costs and all benefits are of the same strength (Howe and Moses, 1999, p. 23); and (b) there is disagreement on what constitutes goodness and happiness, how it will be calculated and who decides. Brooks *et al*. (2014) note that rights-based ethics concern people's liberty, and these trump consequential ethics.

From the consequentialist position, a major ethical dilemma is that which requires researchers to strike a balance between the demands placed on them as professional scientists in pursuit of truth, and the participants' rights and values potentially threatened by the research. This is known as the 'costs/benefits ratio', the essence of which is outlined by Frankfort-Nachmias and Nachmias (1992) in Box 7.1.

Deontological and consequentialist views of ethics concern actions and behaviour. By contrast, a third view is a *virtue ethics* basis which concerns people, and in which one pursues what is good simply because it is what is expected of a good and right person: 'what kind of person we ought to be' (Brooks *et al*., 2014, p. 24). A virtuous person might possess characteristics including loyalty, integrity, respect, sincerity, modesty etc., i.e. virtues. Views of what the virtues are, and what the virtuous person is, may vary by time, place, culture, society etc. For the researcher, Hammersley and Traianou (2012) identify key virtues of dedication, objectivity and independence (respecting academic freedom, professionalism and minimizing negative influences on the research) (pp. 46–51) in the disinterested pursuit of knowledge. Marshall and Rossman (2016, p. 51) note that virtue ethics concern relationships, and researchers have to consider their relationships with participants, stakeholders, sponsors, the research community, individuals, groups, the institution and so on – a complexity of relationships (cf. Ary *et al*., 2002).

These different views of ethics may not sit comfortably together; for example, a utilitarian view might argue that a person who has a healthy body and healthy organs should be killed, and his organs used to save five or six lives of those who otherwise would die, whereas a virtue ethics viewpoint (Hammersley, 2009, p. 213) would argue that this is murder and cannot be justified. Another example of values clashing is where a deontological view might argue that a failing school should be closed, whereas a utilitarian view would argue against its closure because those 2,000 students would go to an even worse school.

A source of tension in considering ethics is that generated by the competing absolutist and relativist positions. The absolutist view holds that clear, set principles should guide the researchers in their work and that these should determine what ought and what ought not to be done (see Box 7.2). To have taken a wholly absolutist stance, for example, in the case of the Stanford Prison Experiment (see Chapter 30), where the researchers studied interpersonal dynamics in a simulated prison, would have meant that the experiment should not have taken place at all or that it should have been terminated well before the sixth day. Indeed Zimbardo (1973), the author of the Stanford Prison Experiment, has stated that the absolutist ethical position, in which it is unjustified to induce any human suffering, would bring about the end of much research, regardless of its possible benefits to

BOX 7.1 THE COSTS/BENEFITS RATIO

The *costs/benefits ratio* is a fundamental concept expressing the primary ethical dilemma in social research. In planning their proposed research, social scientists have to consider the likely social benefits of their endeavours against the personal costs to the individuals taking part. Possible benefits accruing from the research may take the form of crucial findings leading to significant advances in theoretical and applied knowledge. Failure to do the research may cost society the advantages of the research findings and ultimately the opportunity to improve the human condition. The costs to participants may include affronts to dignity, embarrassment, loss of trust in social relations, loss of autonomy and self-determination, and lowered self-esteem. On the other hand, the benefits to participants could take the form of satisfaction in having made a contribution to science and a greater personal understanding of the research area under scrutiny. The process of balancing benefits against possible costs is chiefly a subjective one and not at all easy. There are few or no absolutes and researchers have to make decisions about research content and procedures in accordance with professional and personal values. This costs/benefits *ratio* is the basic dilemma residual in a great deal of social research.

Source: Adapted from Frankfort-Nachmias and Nachmias (1992)

BOX 7.2 ABSOLUTE ETHICAL PRINCIPLES IN SOCIAL RESEARCH

Ethics embody individual and communal codes of conduct based upon a set of explicit or implicit principles and which may be abstract and impersonal or concrete and personal. Ethics can be 'absolute' and 'relative'. When behaviour is guided by absolute ethical standards, a higher-order moral principle is invoked which does not vary with regard to the situation in hand. Such absolutist ethics permit no degree of freedom for ends to justify means or for any beneficial or positive outcomes to justify occasions where the principle is suspended, altered or diluted, i.e. there are no special or extenuating circumstances which can be considered as justifying a departure from, or modification to, the ethical standard.

Source: Adapted from Zimbardo (1984)

society. In other words, ethical absolutism over-constrains research (Hammersley and Traianou, 2012, p. 138); let not the perfect stand in the way of the good or the 'good enough'.

By an absolute principle, the Stanford Prison Experiment must be regarded as unethical because the participants suffered considerably. In absolutist principles a deontological model of research is governed, inter alia, by universal precepts such as justice, honesty and respect. In the utilitarian ethics, ethical research is judged in terms of its consequences, for example, increased knowledge, benefit for many.

Those who hold a relativist position would argue that there can be no absolute guidelines and that ethical considerations will arise from the very nature of the particular research being pursued at the time: the situation determines behaviour. Indeed they would argue that it is essential to respect the context in which the research takes place, culturally, ethnically, socio-economically, and that these should be judged in their own terms (Oliver, 2003, p. 53). This underlines the significance of 'situated ethics' (Simons and Usher, 2000; Hammersley, 2015b), where overall guidelines may offer little help when confronted with a very specific situation. From the standpoint of *situational ethics* (e.g. Simons and Usher, 2000; Oliver, 2003; Hammersley and Traianou, 2012; Hammersley, 2015b), what we should do or what is right to do depends on the situation in question, i.e. judging what to do cannot simply be determined, calculated or logically derived from principles but has to be decided in respect of the presenting situation (i.e. 'bottom-up' rather than 'top-down'): ethical principles *inform* but do not simplistically *determine*. However, this could be challenged on the grounds that it sanctions relativist ethics over absolutist principles.

There are some contexts where neither the absolutist nor the relativist position is clear-cut. Writing of the application of the principle of informed consent with respect to life history studies, Plummer (1983) says:

> Both sides have a weakness. If, for instance, as the absolutists usually insist, there should be informed consent, it may leave relatively privileged groups under-researched (since they will say 'no') and underprivileged groups over-researched (they have nothing to lose and say 'yes' in hope). If the individual conscience is the guide, as the relativists insist, the door is wide open for the unscrupulous – even immoral – researcher.
>
> (Plummer, 1983, p. 141)

He suggests that broad guidelines laid down by professional bodies which afford the researcher room for personal ethical choice offer some way out of the problem. We consider these later in this chapter.

Whilst ethical research concerns principled behaviour, as we will see in this chapter, it is often the case that the research draws on different ethical principles when considering a specific situation, i.e. ethics are situated.

7.3 Sponsored research

Sponsored research does not absolve the researcher from ethical behaviour. For example, it may be considered unethical for the sponsor to control the research or to tell the researcher: (a) how to conduct the research; (b) what results he/she should look for and what findings should be suppressed; (c) what should and should not be reported; (d) to conceal who the sponsor is; and (e) what are the purposes of the research (e.g. Hammersley and Traianou, 2012).

Sponsors may have the right to remain confidential; they may have the right to non-disclosure of who they are and the purposes and findings of the research. Further, researchers will need to consider the effects on the participants of any disclosure of who the sponsors are; for example, if they are told that the research is sponsored by a government office, how will that affect what they say and do?

Sponsors may be oriented towards certain kinds of research, for example that which is related to government policy or which has immediate or direct practical concerns or consequences, and this may suppress 'critical' or sensitive research and prefer user-friendly, policy-affirmative research, or it may exert pressure for a particular style of research to be conducted (e.g. randomized controlled trials) or quantitative research. Sponsors may wish to define the research topic, its limits and how it can or should be done, thereby rendering researchers the instruments of the sponsor. Whilst sponsored research is usually contractual between the researcher and the sponsor, and between the researcher and the participants, and whilst the research may be for the sponsor alone and not for the public, this does not privilege the sponsor in dictating how the research should be conducted and what it should find; in short, 'fixing' the study.

The researcher's responsibilities may lie only in conducting the study and providing the sponsor with a report. What happens to the report after that (e.g. whether it is released completely, selectively or not at all to the public or other parties within the sponsor's organization) is a matter for the sponsor. However, this does not absolve the researcher from decisions about the conduct of the study, and the researcher must retain the right to conduct the study as she or he thinks fit, informed by, but not decided by, the sponsor. The researcher's integrity must be absolute. It is often the case that researchers will negotiate (a) publication rights with the sponsor in advance of the research and (b) what confidentiality the researcher must respect.

The sponsor has a right to expect high-quality, rigorous and useable research. The researcher should not succumb to pressure to:

- betray the confidentiality of the respondents;
- tamper with data, their analysis or presentation to meet a particular objective;
- present selective and unrepresentative data and conclusions;
- make recommendations that do not arise from the data themselves;
- use the data for non-negotiated personal interests, agendas, purposes and advancement;
- conduct a study in which personal research objectives influence the nature, contents and conduct of the research.

The researcher has obligations to the sponsor, but not to doctor or compromise the research. Research is not the same as consultancy in that the researcher may be able to publish from the research (often with due attention to agreed confidentiality and non-traceability etc.) and the researcher owns the research, not the sponsor.

7.4 Regulatory contexts of ethics

Regulatory contexts of ethics have risen massively in influence. There have been moves from guidelines and codes of ethics which advise and inform the professional integrity of researchers and research to regulations, regulatory frameworks and increasing power and legalism to govern, control and police them in a litigious age ('ethics creep') (Haggerty, 2004).

Codes of practice, institutional review boards, university ethics committees, legislation, ethical frameworks and guidelines exist to oversee research in universities and other institutions and can constitute a major hurdle for those planning to undertake research. Ethical codes of the professional bodies and associations as well as the personal ethics of individual researchers are all important regulatory mechanisms. They have a 'gatekeeping' function, to prevent unethical research from taking place, to ensure that no harm comes to participants, and to ensure that due attention has been given by research proposers to the ethical dimensions of their research: for example, risk assessment; acceptable and unacceptable burdens on people and institutions; benefit to different parties; informed consent; confidentiality, non-traceability, privacy, disclosure and protections; control of data; beneficence and non-maleficence; appropriacy of methodology etc.

The celebrated Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979) picks out three key ethical principles that codes of ethics are designed to protect: respect for persons, beneficence and justice (e.g. participants are not denied access to potentially beneficial interventions); as seen below, these are echoed by many organizations. Hammersley and Traianou (2012) identify key principles as: minimization of harm; respect for autonomy (and informed consent); and the protection of privacy (addressing confidentiality and anonymity) (discussed later in this chapter).

Ethical codes of practice are designed to protect the interests of individuals and institutions, to ensure suitably informed consent and to ensure that the proposed research abides by legal requirements and does not violate ethical principles. They address adherence to data protection laws, including that: (a) data will only be used in the way intended by the research and are important for, rather than superfluous to, the research; (b) data are kept securely and only for

as long as necessary; and (c) control and ownership of data are clear, with suitable attention to anonymity (Denscombe, 2014, pp. 319–20).

Professional societies and associations have formulated codes of practice which express the consensus of values within a particular group and which help individual researchers in indicating what is desirable and what is to be avoided. Of course, this does not solve all problems, for there are few absolutes and ethical principles may be open to a wide range of interpretations.

Researchers must take cognizance of ethical codes and regulations governing their practice. Failure to meet these responsibilities on the part of researchers is perceived as undermining the scientific process and may lead to legal and financial penalties and liabilities for individuals and institutions.

Brooks *et al*. (2014) note a distinction between 'foundational principles' and 'practical implications' that have been addressed in regulations, legislation and ethical review boards. The former concerns respect for persons, justice, beneficence and non-maleficence, whilst the latter concerns risk assessment, informed consent, benefits and selection of participants and projects (pp. 27–32). The authors analyse several published research guidelines and codes of practice from research associations, review boards, research sponsors and governments across the world, and note some recurrent ethics matters in them (pp. 30–1), presented here in alphabetical order:

- academic freedom;
- accountability;
- adherence to scientific standards;
- beneficence and non-maleficence (personal and social);
- compliance with the law;
- concern for welfare;
- confidentiality;
- democracy;
- duty of care;
- fairness;
- full information;
- honesty;
- impartiality;
- independence;
- integrity;
- justice;
- objectivity;
- professional competence;
- quality of research;
- reliability;
- respect for persons, including rights, dignity and diversity;
- responsibility (e.g. social, professional, personal) to researchers and others;
- voluntary participation.

Further, in an increasingly information-rich world, it is essential that safeguards be established to protect research information from misuse or abuse. The UK's Data Protection Acts of 1984 and 1998 are designed to achieve such an end. These cover principles of data protection, responsibilities of data users and rights of data subjects. Data held for 'historical and research' purposes are exempted from the principle, which gives individuals the right of access to personal data about themselves, provided the data are not made available in a form which identifies individuals. Such research data may be held indefinitely and their use for research purposes need not be disclosed at the time of data collection, notwithstanding the Freedom of Information Acts which may give the public access rights to data. Personal data (i.e. data that uniquely identify the person supplying them) shall be held only for specified and lawful purposes, and appropriate security measures shall be taken against unauthorized access to, or alteration, disclosure or destruction of, personal data and against accidental loss or destruction of personal data.

Regulatory contexts also include organizations which set legally binding regulations. For example, the United Nations Convention on the Rights of the Child (UNICEF, 1989) sets out fifty-four Articles which are legally binding standards in the fields of, inter alia: non-discrimination; development of full potential; respect for children, their dignity and their views; acting in the interests of the child; protection from harm; active and voluntary participation; and voice. These include statements that place the interests of the child as the primary consideration (Article 3.1), that the child is assured of the right to express his/her own views in matters that affect him/her and that these will be accorded due weight (Article 12.1) and that the child's freedom of expression is protected, including the freedom to 'seek, receive and impart information and ideas' in any media that the child chooses, regardless of frontiers (Article 13.1).

Echoing this, the United Nations Children's Fund (UNICEF) (Graham *et al.*, 2013) produced a wide-ranging document on *Ethical Research Involving Children* which recognizes the cultural location of research (p. 13) and places relationships 'at the core of ethical research' (p. 13). Its key principles are respect, benefit and justice, with their implications for considerations of harm and benefits (non-maleficence and beneficence), informed consent, privacy, confidentiality and payment. Its 'charter' (p. 23) includes seven key statements:

1. Ethics in research involving children is everyone's responsibility.
2. Respecting the dignity of children is core to ethical research.
3. Research involving children must be just and equitable.
4. Ethical research benefits children.
5. Children should never be harmed by their participation in research.
6. Research must always obtain children's informed and ongoing consent.
7. Ethical research requires ongoing reflection.

(Graham *et al.*, 2013, p. 23)

Many institutions of higher education have their own ethics committees, with their own codes of ethics against which they evaluate research proposals. In addition, some important codes of practice and ethical guidelines are published by research associations, for example the British Educational Research Association, the British Psychological Society, the British Sociological Association, the Social Research Association, the American Educational Research Association, the American Psychological Association and the American Sociological Association. We advise readers to consult these in detail.

The British Educational Research Association's (2011) *Ethical Guidelines for Educational Research* are devolved onto:

- *Principles*: 'all educational research shall be conducted within an ethic of respect for: the person; knowledge; democratic values; the quality of educational research; academic freedom' (p. 4).
- *Guidelines*: these address: (a) responsibilities to participants (including sections on: voluntary informed consent; openness and disclosure; rights to withdraw; children, vulnerable young people and vulnerable adults; incentives; detriment arising from participation in research; privacy; disclosure); (b) responsibilities to sponsors of research (including methods and publication); (c) responsibilities to the community of educational researchers (including sections on misconduct and authorship); and (d) responsibilities to educational professionals, policy makers and the general public.

Hammersley and Traianou (2012), writing for the British Educational Research Association, set out five ethical *principles*: minimizing harm; respecting autonomy; protecting privacy; offering reciprocity; and treating people equitably. In light of these principles they raise several issues (pp. 4–12): full and free informed consent; ethical regulation; the gravity of the ethical issues; conflicts between the ethical principles, between different interpretations of them, and how to resolve these differences; having to work with different stakeholders and groups simultaneously; the thrust towards worthwhile research; and situated ethics and judgements (pp. 4–6).

Hammersley and Traianou also include a comprehensive bibliography of many aspects of, and topics in, educational research: diverse perspectives; randomized controlled trials and experimental research; survey research; action research; researching children; Internet research; visual research; narrative and discourse analysis; relationships with funders; anonymity; archiving of data; data protection; philosophical literature on ethics (including feminist ethics and radical approaches); and literature on ethical regulation (with extensive websites). This is an extremely useful reference document, with multiple website references.

The American Educational Research Association's (2011) *Code of Ethics* is organized into principles and ethical standards:

- *Principles*: professional competence; integrity; professional, scientific and scholarly responsibility; respect for people's rights, dignity and diversity; and social responsibility.
- *Ethical standards*: scientific, scholarly and professional standards; competence; use and misuse of expertise; fabrication, falsification and plagiarism; avoiding harm; non-discrimination; non-exploitation; harassment; employment decisions; conflicts of interest; public communications; confidentiality (including privacy and electronic data storage and communication); informed consent (including deception); research planning, implementation and dissemination; authorship credit; publication process; responsibilities of reviewers; teaching, training and administering education programmes; mentoring; supervision; contractual and consulting services; and adherence to the ethical standards of the American Educational Research Association.

These standards run right through the entire research process, from planning to dissemination.

The UK's Economic and Social Research Council's (2015) research ethics framework sets out six key principles:

1. Research participants should take part voluntarily, free from any coercion or undue influence, and their rights, dignity and (when possible) autonomy should be respected and appropriately protected.

2 Research should be worthwhile and provide value that outweighs any risk or harm. Researchers should aim to maximise the benefit of the research and minimise potential risks of harm to participants and researchers. All potential risk and harm should be mitigated by robust precautions.
3 Research staff and participants should be given appropriate information about the purpose, methods and intended uses of the research, what their participation in the research entails and what risks, benefits, if any, are involved.
4 Individual research participant and group preferences regarding anonymity should be respected and participant requirements concerning the confidential nature of information and personal data should be respected.
5 Research should be designed, reviewed and undertaken to ensure recognised standards of integrity are met, and quality and transparency are assured.
6 The independence of research should be clear, and any conflicts of interest or partiality should be explicit.

(Economic and Social Research Council, 2015, p. 4)

One can note here the use of the word 'should', i.e. it has the power of an injunction, a demand (Hammersley and Traianou, 2012, p. 7). The document then provides comprehensive coverage of eleven *minimum requirements* informed by these principles, included in which is the role of Research Ethics Committees to 'protect the dignity, rights and welfare of research participants' (p. 14). The document sets out: (a) a definition of risk and how to approach risk assessment (pp. 27–9); (b) key issues in considering voluntary informed consent (pp. 29–33); and (c) issues to be addressed in an ethics review (pp. 38–9). Further, it carries an extensive list of relevant organizations and their publications on research ethics (pp. 45–9). This is a comprehensive document with useful advice to researchers in considering ethical matters.

It is important to note the references to 'principles' in these documents. Hammersley (2015c) argues that the inclusion of principles (e.g. 'general considerations') (p. 435) is helpful, as: (a) it overcomes the risk that ethics codes and guidelines could become outdated with advances coming on apace in research methodologies and topics; (b) they facilitate agreement among researchers about essential matters; (c) being at a high level of abstraction, they are likely to be inclusive of most research projects (though such abstraction might lend itself to varying interpretation); and (d) they avoid too tight a level of prescription which often encourages researchers 'to follow the letter rather than the spirit of ethical codes' (p. 434).

On the other hand, Hammersley raises the argument that principles may be 'too determinate' and prescriptive in the force of their injunctions and in seeking to derive or deduce from general principles (in a 'quasi-logical way') (2015c, p. 445) those ethical practices which obtain in a specific situation (pp. 443–8). Further, he advances the argument for 'particularism' rather than 'principlism', holding that, since researchers have to focus on specific projects, their decisions are 'situated' and their ethical decisions are taken with reference to the case in hand rather than being derived from sets of principles (p. 441) (see also Hammersley and Traianou, 2012). Whilst principles raise considerations for researchers, they do not necessarily *determine* what ethical decisions should be taken in a specific case; rather, they *inform* such decision making.

The British Psychological Society's *Code of Human Research Ethics* (2014) sets out four principles: respect for the autonomy and dignity of persons (including rights to privacy, self-determination, personal liberty and natural justice) (p. 8); scientific value; social responsibility; and maximizing benefit and minimizing harm. It also states that '[r]esearchers should respect the rights and dignity of participants in their research and the legitimate interests of stakeholders such as funders, institutions, sponsors and society at large' (p. 4). These principles inform its discussions of: risk; valid consent; confidentiality; giving advice; deception; debriefing; ethics review; safeguards for, and working with, vulnerable populations.

With regard to respecting dignity, there is the need to treat participants as equals, not as objects or as subordinate to the researcher. This may mean avoiding treating them as 'subjects' rather than as equals or participants.[1] It also means avoiding stigmatizing groups (e.g. the unemployed; the homeless; religious groups; ethnic groups; those considered deviant by virtue of their sexual orientation, dress, beliefs). The document also indicates that ethical research extends to concerns of: involvement with vulnerable groups; sensitive topics; deception; access to personal or confidential records or sensitive data from third parties (e.g. employers); the potential to induce stress, anxiety, pain humiliation; intrusive interventions that are outside the routine lives of individuals; and that which could lead to negative labelling.

From this very brief excursus into the contents of codes of ethics, it can be seen that they cover similar topics. Ethical codes are a guide, but they cannot dictate to the researcher what to do in a specific, unique situation, nor can they absolve the researcher of responsibility for action taken in the research. The issue here is that ethics are 'situated' (Simons and Usher, 2000):

> while ethics has traditionally been seen as a set of general principles invariantly and validly applied to all situations, … on the contrary, ethical principles are mediated within different research practices and thus take on different significances in relation to those practices.
>
> (Simons and Usher, 2000, p. 1)

The authors state that this implies that situated ethics are 'immune to universalization', because:

> researchers cannot avoid weighing up conflicting considerations and dilemmas which are located in the specificities of the research situation and where there is a need to make ethical decisions but where those decisions cannot be reached by appeal to unambiguous and univalent principles or codes.
>
> (Simons and Usher, 2000, p. 2)

Against ethical codes and regulation

Whilst ethics committees and regulations may be a useful safeguard, for example 'gatekeeping', Farrimond (2013) suggests that ethics committees can be too time-consuming and bureaucratic, unfairly use models from one discipline (e.g. medicine) in another (e.g. social science), are part of a wider 'audit culture', are over-concerned with protecting institutional reputation, make it difficult for researchers to challenge decisions, and lack consistency (pp. 49–51). Brooks *et al*. (2014) report that they can also undermine and discourage professional expertise and reflection. They can be unduly restrictive and lack cultural relevance. In their endeavour to protect even minimal harm from coming to individuals, they may overlook the principle of the greatest good for the greatest number. They may require socio-culturally insensitive or inappropriate ways of ensuring informed consent, for example, requiring such consent to be given on paper-copy pro formas, whereas this is inappropriate in some (e.g. non-western) cultures, and, indeed, they may require individuals to sign such forms when, in reality, in some cultures it is not the individual but the family head, the community and its leaders who should give consent (Lie and Witteveen (2017) comment on the value of filmed informed consent rather than in written form).

Further, ethics committees, codes of practice and regulations may emphasize rule-following in situations where following rules and regulations is insufficient for taking ethical decisions *in situ*. Indeed they may protect the university or institution rather than the research participants, they may prevent important sensitive research from being undertaken, and they may over-simplify complex research situations (cf. Hammersley and Traianou, 2012; Brooks *et al*., 2014, pp. 34–42).

Ethical regulation of research is often conducted by university ethics committees. Whilst the intentions here might be honourable in protecting individuals and institutions, in a blistering paper against ethical regulation by research ethics committees, Hammersley (2009, pp. 212–19) argues that:

a they are incapable of making sound or 'superior' ethical decisions, such that their work will not improve the ethical quality of research. This is because: (i) ethical issues are contentious and there is a lack of consensus among social scientists on ethical matters, principles, priorities of principles and practices, or consequences; (ii) ethical issues and practical research are complex (e.g. secrecy and deception in research); (iii) ethical answers cannot simply be cranked out, mechanistically or algorithmically, but are framed in specific contexts (which may be unknown to ethical committees); (iv) the remit of ethical committees is unclear, for example, whether to approve, prevent, control methodology or topics; (v) ethical committees only need to be persuaded that the researcher has the ethical capability to conduct the research, but this confuses ethical audit with ethical decision making and confuses substance and procedures of ethical review;

b they have no legitimacy or moral superiority/expertise to control researchers, and that this is inherent in ethical principles themselves: (i) researchers should have their autonomy respected; (ii) it is the researchers themselves – and not ethics committees – who have the responsibility for the ethical conduct of research and such responsibility cannot and should not be passed to a committee; (iii) ethics committees must apply the principle of 'informed consent' to researchers, and not just to those being researched; (iv) ethics committees operate prospectively, not only retrospectively, and this kind of prospective regulation is highly unusual in most areas of life; (v) there is almost no evidence that researchers operate unethically apart from some illegal cases, and so the processes of ethics committees are unnecessary, i.e. there is no problem which needs to be fixed;

c they lead to undesirable consequences in research: (i) the bureaucratization of research; (ii) the time and effort required to meet bureaucratic requirements will deter many researchers from proceeding; (iii) research will avoid sensitive, difficult or contested yet important areas and marginalized or powerful groups, i.e. where informed consent may

not be possible; (iv) we may not discover important data; (v) researchers will avoid important research areas because they may consider it difficult to obtain the consent of the ethics committee.

Hammersley argues that ethics committees' roles should be limited to providing advice, providing a forum for discussion on ethical matters and initiating such discussions. This echoes the comment from Howe and Moses (1999, pp. 46–5) that research ethics committees have no special expertise to judge many educational research issues about such-and-such a project, that they are bureaucratic and tend to discharge their duties in a perfunctory manner. They may provide advice and guidance, but not prospective judgements about specific research projects (Howe and Moses, 1999, p. 53).

The difficulty with, and yet the strength of, ethical codes is that they cannot and do not provide specific advice for what to do in specific situations. And the difficulty for ethics committees is that their operations are seen as impractical and, ultimately, anti-ethical, anti-research and anti-researchers. Ultimately, it is researchers themselves, their integrity and their conscience, informed by an acute awareness of ethical issues, underpinned by guideline codes and regulated practice, which should decide what to do in a specific situation, and this should be justified, justifiable, thought through and defensible.

It was observed earlier that many ethical codes and guidelines themselves avoid univalency and unambiguity, arguing, for example, that deception, covert research and the lack of informed consent may be justified. The need for polyvalency (multiple interpretations of what is worthwhile, acceptable and valuable) and situated ethics arises from the practicality of conducting research, the need for sensitivity to socio-political contexts and fairness to disadvantaged groups, and to take account of the diversity and uniqueness of different research practices. What this suggests, then, is that, whilst codes, guidelines and committees may be useful in raising issues and orienting researchers, they cannot decide what should and should not be done in a specific situation; that is for individual researchers and their informed conscience to decide.

7.5 Choice of research topic and research design

Ethical issues enter into considerations of choice of topic and the design and operationalization of the research. The decision on what to research may become a political act, deliberate or not, and the researcher has to consider whose interests are involved in, or at stake in, conducting the research. Why focus on such-and-such in proposing the research? Choice of research topic also raises issues of privacy, sensitivity and access, and we address these later in the chapter.

Here we do not rehearse the issues of 'partisan research' that we addressed in Chapter 3, though readers may find it useful to review that chapter. Rather, we address issues of the kind of research proposed. Researchers may find themselves having to defend not only their research design but their methodology in the face of, for example, questions from ethics committees or sponsors who lean towards quantitative methods and experimental approaches and away from what they perceive to be 'soft', unscientific qualitative research. This may extend to their being uncomfortable with covert research or research which, as it unfolds over time or is conducted in non-mainstream or 'different' cultures, may raise questions of informed consent or where the research will go (e.g. 'blue-skies' research) and what the research will 'deliver' (e.g. issues which routinely face ethnographic research).

The issue goes further, into instrumentation: some sponsors may not be friendly towards survey research (Brooks *et al*., 2014), particularly if it might find 'unwelcome data' (p. 72); they may either bar researchers from identifying individuals or, by contrast, positively require parties and individuals to be identified. As discussed earlier, this is illegitimate, as sponsors cannot dictate to researchers how to go about their business.

Even apparently innocuous instruments such as questionnaires broach ethical issues, as respondents typically have to answer in terms of categories already decided by the researcher, and this risks reducing the participants to data objects rather than agentic people (Hammersley and Traianou, 2012c, p. 12). Observation, participant or non-participant, intrudes into people's lives and privacy.

Researchers will also need to consider where their research will take place, for example, in classrooms, off-site (e.g. for confidentiality and 'neutral territory'), as context often influences the research. For example, interviewing children in school may be more comfortable for children than outside school or, indeed, the opposite may be the case (Morrison, 2013a). Conducting interviews in school may make students feel obliged to participate when in fact they would prefer not to participate, or might discourage them from being honest in favour of saying or doing what they think the school would require. Schools as hierarchical institutions may exert an influence on students (Brooks *et al*., 2014, p. 76).

In setting research agendas, the researcher may also face dilemmas in deciding whose agendas to serve: the insiders' (the participants, the sponsors) or the outsiders' (the researcher, the reader, the audiences of the research, the public). Whilst this may lead to some negotiation of research agendas and methodologies (e.g. in sensitive research, see Chapter 13), this is an ethical issue and the integrity of the researcher cannot be sacrificed to insider power. On the one hand it may be better not to do the research than to compromise ethical principles in deciding the agenda and the conduct, methodology and reporting of the research, how the research will be conducted and reported, and to whom. On the other hand it may be better to give ground in some quarters in order to gain ground in others: better to do some research rather than no research at all or better to make some matters public rather than no matters at all. This returns to the issue of 'situated ethics' introduced earlier: researchers have to decide how to behave ethically in each specific situation.

In identifying the research topic and the design of the research, the researcher moves to planning the conduct of the research, and this raises a huge range of issues, which we consider below. These ethical concerns need to be addressed very early on in the planning, design and conduct of the research, including: informed consent; non-maleficence; beneficence; human dignity; privacy; anonymity; confidentiality; betrayal and deception; Internet ethics; ethics and evaluative research.

Ethics and the quality of the research

The research design, and indeed the research itself, have an ethical duty to demonstrate quality. Put simply, badly designed research is a breach of ethics (Farrimond, 2013, p. 75). As Hammersley and Traianou (2012) argue, research ethics is not only about treating people correctly, i.e. procedural matters such as rights, interests and duties, but about the 'primary obligation' (p. 1) to answer worthwhile questions; to pursue, produce, test and defend valid factual knowledge (e.g. descriptions, explanations, interpretations, conclusions, theories etc.) on the basis of full evidence; and to make a significant, relevant contribution to knowledge. The fundamental purpose of research is the production of valid, relevant, worthwhile and significant knowledge. It does not have as a *necessary* goal the practicality or political attractiveness of such knowledge; there may be motives for undertaking the research but these are not the purposes of the research itself (p. 134).

Producing knowledge includes, inter alia: a rigorous and coherent research design that demonstrates fitness for purpose; appropriate sampling, methodology and instrumentation; transparency, usefulness and validity; scholarly and scientific merit; significance and advancement of the field (e.g. substantively, conceptually, methodologically); scientific value; risk assessment and minimization; and transparency. Research which does not advance the field or which is of poor quality may be a waste of time and resources, and may be literally useless, all of which violate ethics.

Ethics in educational research also affect those not directly involved in the study but who may be affected by it, for example, children and parents. Ensuring the quality of research is an ethical matter, as poor-quality educational research can cause harm, particularly if used for decision making, for example, by funding bodies, parents, governments and policy makers. In other disciplines such as medicine, poor-quality research would not pass muster and would be prevented from taking place (and this might be a legitimate role for ethics committees). In poor-quality research: the evidence is weak; it uses inappropriate designs, methods and data analysis; conclusions made do not follow from the evidence; no warrants are made to link evidence and conclusions (Gorard, 2013); reporting is biased (we discuss bias in data analysis, reporting and dissemination later in this chapter); and claims made from the research are not supported by the research and data provided. Indeed taxpayers' money is wasted.

In poor-quality research, claims concerning 'impact' are made from inappropriate sampling, lack of baseline data for enabling comparisons to be made, attrition and missing data, lack of counterfactual analysis and data, and inadequate consideration of alternative explanations of findings. Ethical researchers have a duty of care to ensure that their research is of the highest quality.

Robson (1993, p. 33) raises ten questionable practices in social research:

- involving people without their knowledge or consent;
- coercing them to participate;
- withholding information about the true nature of the research;
- otherwise deceiving participants;
- inducing them to commit acts diminishing their self-esteem;
- violating rights of self-determination (e.g. in studies seeking to promote individual change);
- exposing participants to physical or mental stress;
- invading their privacy;
- withholding benefits from some participants (e.g. in comparison groups);
- not treating participants fairly, or with consideration, or with respect.

He calls these 'questionable practices' rather than areas to be proscribed, and this indicates that they are not black and white, right or wrong matters. They constitute ethical dilemmas for the researcher.

Earlier this chapter introduced the consequentialist costs/benefits ratio. Frankfort-Nachmias and Nachmias (1992) express this as a conflict between two rights: the rights to conduct research in order to gain knowledge versus the rights of participants to self-determination, privacy and dignity. This constitutes a fundamental ethical dilemma of the social scientist for whom there are no absolute right or wrong answers. Which proposition is favoured, or how a balance between the two is struck, will depend on the background, experience and values of the individual researcher.

7.6 Informed consent

The principle of informed consent concerns autonomy, and it arises from the participant's right to freedom and self-determination. Being free is a condition of living in a democracy, and when restrictions and limitations are placed on that freedom they must be justified and consented to, as in research. Consent thus protects and respects the right of self-determination and places some of the responsibility on the participant should anything go wrong in the research. Self-determination requires participants to have the right to weigh up the risks and benefits of being involved in a piece of research, and deciding for themselves whether to take part (Howe and Moses, 1999, p. 24). As part of the right to self-determination, the person has the right to refuse to take part, or to withdraw once the research has begun (see Frankfort-Nachmias and Nachmias, 1992). Thus informed consent implies informed refusal.

Informed consent has been defined by Diener and Crandall (1978) as those procedures for individuals to choose whether or not to participate in the research, once they have been told what it is about and what it requires, i.e. all those factors which might influence their decision (p. 57). This definition involves four elements: competence, voluntarism, full information and comprehension.

'Competence' implies that responsible, mature individuals will make correct decisions if they are given the relevant information. It is incumbent on researchers to ensure they do not engage individuals incapable of making such decisions because of either immaturity or some form of impairment. The United Nations Convention on the Rights of the Child (1989) and Graham *et al*. (2013) underline the importance of involving children in decisions that may affect them, and this extends to them giving informed consent provided that they are competent to understand what is involved in the research, and in the UK it means even if this overrides their parents' wishes or if children are below their biological age for assuming maturity (Brooks *et al*., 2014, pp. 82–7).

'Voluntarism' entails ensuring that participants freely choose to take part (or not) in the research and guarantees that exposure to risks is undertaken knowingly and voluntarily.

'Full information' implies that consent is fully informed, though in practice it is often impossible or even undesirable for researchers to inform participants on everything (see section below on 'Deception') and, as we see below, on those occasions when the researchers themselves do not know everything about the investigation and how it will unfold. In such circumstances, the strategy of 'reasonably informed consent' has to be applied. Box 7.3 illustrates a classic set of guidelines used in the United States that are based on the idea of reasonably informed consent (Department of Health, Education and Welfare, 1971).

'Comprehension' refers to the fact that participants fully understand the nature of the research project, even when procedures are complicated and entail risks.

If these four elements – competence, voluntarism, full information and comprehension – are present, participants' rights will have been given appropriate

BOX 7.3 GUIDELINES FOR REASONABLY INFORMED CONSENT

1. A fair explanation of the procedures to be followed and their purposes.
2. A description of the attendant discomforts and risks reasonably to be expected.
3. A description of the benefits reasonably to be expected.
4. A disclosure of appropriate alternative procedures that might be advantageous to the participants.
5. An offer to answer any inquiries concerning the procedures.
6. An instruction that the person is free to withdraw consent and to discontinue participation in the project at any time without prejudice to the participant.

Source: US Department of Health, Education and Welfare (1971)

consideration. This also raises questions of who is the appropriate party to give informed consent, for example, an individual, a community, an institutional head, to whom such consent applies, and consent for what ('freedom from' and 'freedom for') (Hammersley and Traianou, 2012, p. 80). Further, such informed individual consent may not be a feature of, say, oppressive regimes in which participation may be mandated; does this mean that research cannot take place here? Do such people have the right not to be researched?

Further, Frankfort-Nachmias and Nachmias (1992) note that informed consent may not always be necessary (e.g. deception may be justified), but that, as a general rule, the greater the risk, the more important it is to gain informed consent. More widely, it raises the question of whether informed consent is really required and, if so, what form it should take (cf. Hammersley and Traianou, 2012, chapter 4).

Informed consent is one of the most problematic issues in educational research, as it raises a lengthy list of concerns, for example:

- Should consent be an individual, family, institutional or communitarian decision?
- Who gives consent, and for whom, for what and for how long (e.g. longevity of data storage)?
- What constitutes 'consent'?
- Who is competent to give consent, and on whose behalf?
- Can children override parents' wishes?
- What pressure (deliberate or not) on people and institutions is there to give consent?
- What does 'voluntary' really mean in 'voluntary consent'?
- In whose interests is consent given or withheld?
- How is consent given in different cultures?
- How to protect vulnerable people in giving consent.
- What degree of informality and formality is appropriate in consent giving?
- What are the possible consequences (and to whom) of consent or non-consent?
- How do power differentials affect consent giving?
- Is biological age of consent 'good enough' for giving consent?
- What are the relationships between consent and confidentiality?
- How much information is it necessary to give or withhold from participants when asking for informed consent (what does 'fully informed' mean and require)?
- Should incentives be offered to gain consent?
- How can consent be addressed in covert research?
- What tensions arise in considering consent and action research (where the researcher is the powerful teacher)?
- Is deception justified?
- How can consent be given when what happens may not be fully known in advance of the research (e.g. in exploratory research)?
- How can consent be addressed in online research?

What starts out as a simple label – 'informed consent' – raises an enormous list of concerns, and we address these in the pages below.

Whilst some cultures may not be stringent about informed consent, in others there are strict protocols for informed consent. What form should or does consent giving take? In some cultures, consent has to be given in writing; in others, such written consent is deemed to be suspicious, threatening, insulting or culturally inappropriate (cf. Hammersley and Traianou, 2012, p. 89; Farrimond, 2013, pp. 33–4), and (unrecorded) oral consent, or even a nod of a head by the appropriate person, is sufficient. Brooks *et al*. (2014) comment that there is a risk that seeking written informed consent, particularly from individuals, is 'fundamentally western and masculinist' (p. 83) and neglects communitarian requirements for giving consent, as, in some cultures, it is the community which is the gatekeeper, not the individual (cf. Howe and Moses, 1999, pp. 33–4). In other words, consent is culturally situated (Marshall and Rossman, 2016, p. 55), and the giving of informed consent differs in individualist and collectivist cultures (p. 57). Written consent might be seen as bringing a level of formality into what some cultures and communities would prefer to keep on an informal footing (Crow *et al*., 2006, pp. 88–9).

Informed consent, aver Howe and Moses (1999), is a cornerstone of ethical behaviour, as it respects the right of individuals to exert control over their lives and to take decisions for themselves. How far this extends to, for example, parents, counsellors, groups and communities is not a black-and-white matter. What happens, for example, if the researcher wishes to study child abuse at home; does she need the parents' permission? Is not the requirement for parental consent for research on children simply being too adult-centred (Brooks *et al*., 2014, p. 158)?

Informed consent often concerns access (Hammersley and Traianou, 2012), for example, to people, documents, institutions, settings and information. This, in turn, requires attention to how to secure consent and from whom (whose consent is required), for what (e.g. information, purposes and for what subsequent uses), for whom (on whose behalf and covering which

people: participants, gatekeepers, others), for how long (pp. 82–90), and how to give information (in what form or medium, and with how much formality/informality) (p. 96). It concerns what counts as 'free', under what constraints and persuasions (pp. 91–2) and whether consent is actually necessary, extending to issues of what constitutes public and private information, places and settings and how – or whether – consent relates to covert research.

Ruane (2005, p. 21) raises the question of 'how much information is enough?'. She argues that this may be an unknown, not necessarily deliberately withheld. Is it justifiable to give 'partial truths' in providing information (Hammersley and Traianou, 2012, p. 94)? Further, just as providing information may bias the results (i.e. it is important for the integrity of the research *not* to disclose its purposes or contents, e.g. the Milgram experiments, see Chapter 30), it may actually confuse the respondents. But if the researcher decides that partial information is preferable, does not this violate the principle of informed consent; is it being dishonest?

Educational research which involves children must recognize that they may not be on equal terms with the researcher (e.g. in terms of power) and it is important to keep this in mind at all stages in the research process, including the point where informed consent is sought. In this connection we refer to the important work from UNICEF (Graham *et al*., 2013) and the codes of practice introduced earlier.

There are other aspects of the problem of informed consent (or refusal) in relation to young or very young children (Greig and Taylor, 1999, pp. 143–55), not least of which is the need to abide by the requirements of legislation on working with children and on child protection. Greig and Taylor argue (p. 150) that non-therapeutic research should only be conducted with children where there is negligible risk and where the informed consent of gatekeepers (e.g. guardians and parents) has been obtained in advance, including how data will be stored (e.g. ICT-related issues), destroyed (e.g. confidential data or audio/visual recordings) upon completion of the research, how recording data may be switched off during an interview and how data will be used. For a fuller guide on ethical issues in conducting research on early childhood education, see Mukherji and Albon (2010) and the UNICEF document, *Ethical Research Involving Children* (Graham *et al.*, 2013).

Gatekeepers are in a very responsible position and they should not be overlooked. Oliver (2003, p. 39) comments that they have much more at stake – to lose – than researchers, since, whereas researchers can move on from one participant or researcher field to another, gatekeepers live with the daily consequences of the research and its effects on participants. Researchers may have an ethical obligation to seek the informed consent of gatekeepers. In turn, it must be recognized that gatekeepers also consider their own interests – protecting or promoting them – and hence may try to steer the research in certain directions, block or steer access, or try to control the dissemination of the results (Hammersley and Traianou, 2012, p. 50).

Seeking informed consent with regard to minors involves two stages. First, researchers consult and seek permission from those adults responsible for the prospective minors, and second, they approach the young people themselves. The adults in question will be, for example, parents, teachers, tutors, counsellors, youth leaders, or team coaches, depending on the research context. The point of the research will be explained, questions invited and permission to proceed to the next stage sought. Objections, for whatever reason, will be duly respected.

While seeking children's permission and cooperation is an automatic part of some research (e.g. a child cannot unknowingly complete a simple questionnaire), the importance of informed consent in some research is not always recognized. Speaking of participant observation, for example, Fine and Sandstrom (1988) say that researchers must provide a credible and meaningful explanation of their research intentions, especially in situations where they have little authority, and that children must be given a real and legitimate opportunity to say that they do not want to take part (cf. Graham *et al.*, 2013). Where participants do refuse, they should not be questioned, their actions should not be recorded and they should not be included in any book or article (even under a pseudonym). Where they form part of a group, they may be included as part of a collectivity. Fine and Sandstrom (1988) consider that such rejections are sometimes a result of mistrust of the researcher. They suggest that at a later date, when the researcher has been able to establish greater rapport with the group, those who refused initially may be approached again, perhaps in private.

Two particular groups of children require special mention: very young children and those not capable of making a decision. Researchers intending to work with pre-school or nursery children may dismiss the idea of seeking informed consent from their would-be participants because of their age, but Fine and Sandstrom (1988) and UNICEF (Graham *et al.*, 2013) would recommend otherwise. Even though such children might not understand what research is, the authors advise that the children be given some explanation. For example,

an explanation to the effect that an adult will be watching and playing with them might be sufficient to provide a measure of informed consent consistent with the children's understanding. Fine and Sandstrom (1988) and Graham *et al*. (2013) comment that children should be told as much as possible, and that steps should be taken to ensure that they understand, and that this obtains regardless of their age.

The second group consists of those children in a research project who may not meet Diener's and Crandall's (1978) criterion of 'competence' (a group of psychologically impaired children, for example – the issue of 'advocacy' applies here). In such circumstances there may be institutional or local authority or legal guidelines to follow. In the absence of these, the requirements of informed consent would be met by obtaining the permission of those acting *in loco parentis* (e.g. headteachers) or who have had delegated to them the responsibility for providing informed consent by the parents.

Two cautions: first, where an extreme form of research is planned, parents would have to be fully informed in advance and their consent obtained; and second, whatever the nature of the research and whoever is involved, should a child show signs of discomfort or stress, the research should be terminated immediately. For further discussion on the care that needs to be exercised in researching with children, we refer readers to Greig and Taylor (1999), Holmes (1998), Graue and Walsh (1998) and UNICEF (Graham *et al.*, 2013).

Informed consent applies not only to children, but to a range of vulnerable groups, for example, adults, the disabled, those who cannot speak, see or hear, those in hospital, those in care, those suffering from autism (cf. Coch, 2007; Waltz, 2007; Brooks *et al*., 2014). Oliver (2003, pp. 35–6) defines vulnerable groups as those people or categories of people who, for whatever reason, may not have sufficient understanding to be able to give informed consent to the research. In many cases ethics committees will require a full indication of how the ethics of the research will be addressed here (Crow *et al.*, 2006, p. 86).

Informed consent requires an explanation and description of several factors, including:

- the purposes, contents, procedures, reporting and dissemination of the research;
- any foreseeable risks and negative outcomes, discomfort or consequences and how they will be handled;
- benefits that might derive from the research;
- incentives to participate and rewards from participating;
- right to voluntary non-participation, withdrawal and re-joining the project;
- rights and obligations to confidentiality and non-disclosure of the research, participants and outcomes;
- disclosure of any alternative procedures that may be advantageous;
- opportunities for participants to ask questions about any aspect of the research;
- signed contracts for participation.

Brooks *et al*. (2014, p. 94) suggest that consideration can also be given to:

- the sponsors of/source of funding for the research;
- why the participants have been approached;
- how anonymity is assured;
- how data will be reported;
- contact details of the researcher(s).

Researchers who seek informed consent must ensure – check – that participants really do understand the implications of the research, not mindlessly sign a consent form. They may need time to digest the information given before consenting or not consenting. Researchers will need to decide what to include in informed consent, not least of which is the issue of volunteering. Participants may feel coerced or pressurized to volunteer (e.g. by a school principal), or may not wish to offend a researcher by refusing to participate, or may succumb to peer pressure to volunteer (or not to volunteer), or may wish to volunteer for reasons other than the researcher's (e.g. to malign a school principal or senior colleagues, to gain resources for his or her department, to gain approval from colleagues).

For example, in action research, the researcher is often the child's own teacher. Will the child really be given the right not to take part, or is the action research seen less like research and more like the carrying out of a professional duty to ensure that the best possible education is being promoted, i.e. part of the normal practice of improving curricula, teaching and learning, and hence not requiring the consent of the child or the parents?

It is important to ensure that participants are not 'railroaded' into participating, for example, by a school principal who makes the decision for the staff, or where staff are not given sufficient time to come to a decision on whether or not to participate, or where staff do not wish to appear unhelpful to researchers (who, indeed, may be friends or acquaintances of the researcher), even though they actually would rather not take part in

the research (Oliver, 2003, p. 27). The choice of whether or not to participate must be genuinely free, with no negative repercussions for not taking part, and no feelings of researchers having taken advantage of powerless participants (cf. Graham *et al.*, 2013).

Arguments against informed consent

There are some research methods where it is impossible to seek informed consent. Covert observation, for example, as used in Patrick's (1973) study of a Glasgow gang (Chapter 15), or experimental techniques involving deception, as in Milgram's obedience-to-authority experiments (Chapter 30), would, by their very nature, rule out the option. And, of course, there may be occasions when problems arise even though consent has been obtained. Burgess (1989), for example, cites his own research in which teachers had been informed that research was taking place but in which it was not possible to specify exactly what data would be collected or how they would be used. It could be said, in this particular case, that individuals were not fully informed, that consent had not been obtained and that privacy had been violated.

Some researchers advocate informed consent on the grounds that it yields better data because it is a consequence of establishing rapport and trust between researchers and participants (e.g. Crow *et al*., 2006, p. 76). Indeed it might bring better participation rates in research, as participants might be more likely to agree to being involved if they are given the 'full picture' of the research or if assurances of confidentiality are given. On the other hand, informed consent is seen less positively, as it renders some research (e.g. necessarily covert research) unresearchable, and it provides poorer participation rates where some participants may be reluctant to sign a consent form or may regard the research as too bureaucratic (p. 88), antagonistic, coercive and alienating.

Informing people of the research might provoke the Hawthorne effect (see Chapter 14) or might disturb the natural behaviour of participants (Oliver, 2003, p. 53) as they will be conscious of being watched. Hence full disclosure of the research aims and purposes might distort the research process. Whether this amounts to deception is addressed below (see section on 'Deception').

Seeking formal informed consent might lead to a narrow range of data and a neglect of the richest, most authentic data, as participants might become more guarded in what they disclose (e.g. about relationships). Indeed in some cultures, Oliver (2003, p. 103) writes, participants may find it an unusual experience to be asked to complete a questionnaire, and they may regard it as a 'test'. Howe and Moses (1999) ask how realistic it is to obtain informed consent from both parents to interview their child if the parents are separated (p. 89). The effects of all of these difficulties might lead to research only concerning itself with 'safe', easily researchable topics and to the neglect of research into vulnerable and excluded groups. By contrast, Humphreys (1975), the author of the celebrated study *Tearoom Trade* (1970), a study of homosexual meeting arrangements, wrote in his 1975 postscript on ethics that 'the greatest harm a social scientist could do to this man would be to ignore him' (p. 169).

Informed consent may not be possible in covert research, or research in which important yet sensitive issues or groups are being investigated (see Chapter 13), and it is only through covert research and perhaps deception that one can gain access to such sensitive groups or practices. It might be important to research such groups (see Mitchell's (1993) defence of secrecy in research for the public good). Similarly, in ethnographic research, the researchers may not know in advance what kind of data will be collected, from whom and how. In these circumstances Brooks *et al*. (2014) note that an 'ethics of care' might be more suitable than 'informed consent' (p. 89).

Wax (1982, p. 44) argues that informed consent offers both 'too much and too little': 'too much' in the sense that it is 'overscrupulous and disruptive', particularly in emergent situations and qualitative research where casual conversations figure highly as field notes, and 'too little' in the sense that field researchers often require much more than informed consent, for example, they seek trust, 'active assistance' from participants and 'colleagueship'. Indeed he suggests that informed consent reinforces asymmetries of power between researchers and participants, rather than equalizing them.

If the research involves participants in a failure experience, isolation or loss of self-esteem, for example, researchers must ensure that the participants do not leave the situation more humiliated, insecure, alienated and worse off than when they arrived. From the participant's point of view, procedures which involve loss of dignity or injury to self-esteem, or affect trust in rational authority, are probably most harmful in the long run and may require the most carefully organized ways of recompensing the participants if the researcher chooses to carry on with those methods.

With particularly sensitive areas, participants need to be fully informed of the dangers of serious after-effects. There is reason to believe that at least some of the obedient participants in Milgram's (1963) experiments (see Chapter 30) came away from the experience

with lower self-esteem, having to live with the realization that they were willing to yield to destructive authority to the point of inflicting extreme pain on a fellow human being (Kelman, 1967). Researchers may need to reflect attitudes of compassion, respect, gratitude and common sense without being too effusive. Participants clearly have a right to expect that the researchers with whom they are interacting have some concern for their (participants') welfare.

As a general rule, informed consent is an important principle, but, as noted above, it may not always be fully or easily applied, or desirable.

7.7 Non-maleficence, beneficence and human dignity

Non-maleficence (do not harm) is enshrined in the Hippocratic oath, in which the principle of *primum non nocere* (first of all, do no harm) is held as a guiding precept; so also with educational research. Adopting consequentialist ethics, the research should not damage the participants physically, psychologically, emotionally, professionally, personally and so on. For example, participants may find it very distressing to relive the experience of being bullied by students or other staff and researchers must decide whether or how to proceed here, with due attention to informed consent and right not to take part (cf. Oliver, 2003, p. 32).

Non-maleficence requires researchers and participants to consider carefully the possible consequences of the research on participants and the research (e.g. the negative effects on the participants and the researchers). Hammersley and Traianou (2012) argue that all research, just as everyday life, involves the risk of harm, it cannot be removed completely so the task of the researcher is to minimize it (p. 57). Further, what constitutes harm or the level of risk (small to significant) is a matter of judgement (p. 57).

Non-maleficence considers the need to avoid doing harm to participants. At first sight this seems uncontentious: of course we do not wish to bring harm to our research participants, and it is a golden rule that the research must ensure that participants are no worse off at the end of the research than they were at the start of the research. However, what constitutes 'harm' is unclear; one person's harm may be a society's benefit, and whether a little harm for a few is tolerable in the interests of a major benefit for all, or even for the person concerned, throws into relief the tension involved here.

Issues of harm and risk also raise the question of what constitutes 'worthwhile' knowledge: is the benefit worth the risk or the harm? For example, having participants recalling distressing or traumatic experiences at interview may turn out to be beneficial for them; it may be therapeutic in coming to terms with them. Harm and risk involve matters of judgement, and these involve carefully weighing complex issues and the potential degree of harm. In open-ended research or research in naturalistic settings which may be uncontrollable, for example, certain types of qualitative research, this may be unpredictable, and indeed the contexts themselves may be involving harm (Hammersley and Traianou, 2012, p. 65).

Hammersley and Traianou (2012) also identify several kinds of potential harm: (a) pain or physical injury; (b) psychological or emotional damage; (c) material damage or loss; (d) reputational or status damage or loss; (e) damage to an activity or project in which participants are involved (p. 62). Attention must be given not only to the immediate harm in (a) to (e) here, but also to 'knock-on' effects and their duration, for example, on quality of life. Attention has to be given not only to the potential degree of harm but to the degree of severity of its consequences (pp. 63–4). The authors note, for example, that visual research with children (even giving the children themselves the video cameras), in which they or their circle of contacts can be identified on video or film, in photographs or on Internet sites, poses potential risks of harm to participants or to others identified in the visual data, raising issues of whose informed consent is needed in the research, who is responsible for what and what can or cannot be foreseen (pp. 69–71).

The question here is whether the end justifies the means. It involves the 'dirty hands' dilemma: whether causing a small harm to a few is justified in terms of bringing about a greater good, for example, to society and the public good. As a general principle we advocate the application of *primum non nocere*, and, indeed, ethics regulatory boards are guided heavily by this principle. However, there could be tensions here. What do you do if you discover that the headteacher has a serious alcohol problem or is having an illicit affair with a parent? What do you do if your research shows that your teacher friend in a school has very serious weaknesses, such that their contract should be terminated in the interests of the students? Harm may also accrue to individual participants, but it does not rest there; it can extend to the researcher(s), institutions, groups and communities being researched, the researcher's own workplace and colleagues, publishers and those with whom the researcher may not have had direct personal contact.

When researchers are confronted with dilemmas such as these (however few they may be), it is generally considered that they resolve them in a manner that

avoids the extremes of, on the one hand, giving up the idea of research and, on the other, ignoring the rights of the participants. At all times, the welfare of participants should be kept in mind (cf. British Educational Research Association, 2011), even if it involves compromising the impact of the research. In the final reckoning, the decision to go ahead with a research project rests on a subjective evaluation of the costs both to the individual and society.

Bailey (1994, p. 457) suggests that there are several approaches that can be used to avoid harming research participants, including:

- using computer simulations;
- finding a situation in which the negative effects of harm already exist, i.e. where the research does not have the responsibility for having produced these conditions;
- applying only a very low level of potential harm, or for only a short period of time, so that any effects are minimal;
- informed consent (providing details of the potential negative effects and securing participants' consent);
- justifying the research on the grounds that the small amount of harm caused is much less than the harm caused by the existing situation (which the research is trying to improve);
- using samples rather than complete populations, so that fewer people are exposed to the harm;
- maintaining the privacy of participants through the use of aggregated or anonymized data.

Whilst some of these are uncontentious, others in this list are debatable, and researchers will need to be able to justify the decisions that they reach.

The complement of non-maleficence is beneficence: what benefits will the research bring, and to whom, and how? Many would-be participants could be persuaded to take part in research if it is made clear that it will, or may, bring personal, educational and social benefits. For example, it may lead to the improvement of learning, increased funding and resources for a particular curriculum area, improved approaches to the teaching of a subject, increased self-esteem for students, additional teachers in a school, increased self-awareness in the participants (Oliver, 2003, p. 35) and so on. Whilst it is sometimes worth including a statement of potential benefit when contacting schools and individuals, it may also be an actual requirement for ethics regulatory boards or sponsors.

Although it may be fanciful to believe that a single piece of research will automatically lead to improvement, the ethical question raised here – who benefits? – suggests that a selfish approach to the benefits of the research by the researcher is unethical. This point requires researchers to do more than pay lip service to the notion of treating research participants as subjects rather than as objects to be used instrumentally – research fodder, so to speak – imbuing them with self-esteem and respect.[1] One can treat people with respect but still the research may make no material difference to their lives.

Whilst it is impossible to argue against treating people with dignity and respect, it also raises the issue of the obligations and commitments of the researcher. Let us say that the researcher has been working closely in a school for one or two years; surely that researcher has an obligation to improve the lives of those being researched, rather than simply gathering data instrumentally? To do the latter would offend reciprocity (see section below on 'Reciprocity'). The issue is tension-ridden: is the research *for* people and issues or *about* people and issues? We have to be clear about our answer to the question 'what will this research do for the participants and the wider community, not just for the researcher?'

Beneficence, whilst eminently worthy, may not be the main purpose of educational research. Hammersley and Traianou (2012) make a powerful case for regarding educational knowledge as the production of valid and significant knowledge; if, in so doing, it brings beneficence then this may be a bonus, a consequence, not a purpose.

7.8 Privacy

With increasing surveillance in everyday life, with electronic storage and retrieval, privacy and its protection become a difficult and contested terrain. Further, there may be some activities or places which, by their very nature, are intensely private and to intrude into them is to break taboos. Qualitative research, in particular, has considerable potential to invade privacy. Privacy touches all aspects of the research enterprise: choice of topic, research design, foci, participants, instrumentation, questions asked, data and their collection, data analysis, reporting and dissemination.

Privacy is a primordial value, a 'basic human need' (Caplan, 1982, p. 320), which, like the right to self-determination, 'trumps' utilitarian calculations (Howe and Moses, 1999, p. 24). Its corollaries are anonymity, confidentiality and informed consent. It has been considered from three different perspectives by Diener and Crandall (1978). These are: the sensitivity of the information being given; the setting being observed; and dissemination of information. Sensitivity of

information refers to how personal or potentially threatening is the information collected by the researcher. Certain kinds of information are more personal than others and may be more threatening, for example, religion, ethnicity, sexual practices, income, values and other personal attributes such as intelligence, honesty and courage may be more sensitive items than name or age. Thus, the greater the sensitivity of the information, the stronger must be the safeguards called for to protect the privacy of the participants. These vary by culture: in one culture what is private or sensitive, what is legitimate public territory and who can have access to private matters may not be so in another, and the researcher needs to judge this. Similarly, what can and should be published (a central feature of academic research) or kept private may be contestable.

The term 'private' has different meanings (Hammersley and Traianou, 2012, pp. 111–12), for example: a 'home area' or territory (e.g. a street, a district); 'not publicly owned' (but these may still be open to the public, e.g. a shopping mall); restricted areas (who can and cannot enter them, e.g. one's home), be they publicly owned (e.g. a local authority school) or privately owned (a private school); a private activity which takes place in a public place. These, the authors note (p. 113), can be used to judge not only whether something is private, but the degree of privacy involved and the legitimacy of observing or collecting data about them with and without consent, and with varying degrees of intrusion. Indeed what counts as 'intrusion' and legitimate or illegitimate intrusion is also contestable. Private matters will affect what, how, when and where questions are asked (p. 114).

The setting being observed may vary from very private to completely public. The home, for example, is considered one of the most private settings and intrusion into people's homes without their consent is forbidden by law. As is the case with most rights, privacy can be voluntarily relinquished. Participants may choose to give up their right to privacy by either allowing a researcher access to sensitive topics or settings or by agreeing that the research report may identify them by name. The latter case at least would be an occasion where informed consent would need to be sought.

Generally speaking, if researchers intend to probe into the private aspects or affairs of individuals, their intentions should be made clear and explicit and informed consent should be sought from those who are to be observed or scrutinized in private contexts. Other methods to protect participants are anonymity and confidentiality and we examine these below.

Privacy is more than simple confidentiality (discussed below). The right to privacy means that a person has the right not to take part in the research, not to answer questions, not to be interviewed, not to have their home intruded into, not to answer telephones or emails, and to engage in private behaviour in their own private place without fear of being observed. It is *freedom from* as well as *freedom for*. This is frequently an issue with intrusive journalism. Hence researchers may have an obligation to inform participants of their rights to refuse to take part in any or all of the research, to obtain permission to conduct the research, to limit the time needed for participation and to limit the observation to public behaviour.

Individual 'right to privacy' is usually contrasted with the public's 'right to know' (Pring, 1984, 2015): what is in the public interest and serves the public good versus the individual's right to privacy. Researchers will need to decide this on a case-by-case basis: 'situated ethics'.

Chapter 8 discusses the threats to privacy introduced by online research, and we recommend researchers to consult this chapter.

7.9 Anonymity

One way of addressing privacy and protection from harm is by anonymity. Frankfort-Nachmias and Nachmias (1992) underline the need for confidentiality of participants' identities, holding that any violations of this should be made with the agreement of the participants. The essence of anonymity is that information provided by participants should in no way reveal their identity. The obverse of this is personal data which uniquely identify their supplier. A participant is considered anonymous when the researcher or another person cannot identify the participant from the information provided. For example a questionnaire might only contain a number instead of a person's name. Where this situation holds, a participant's privacy is guaranteed, no matter how personal or sensitive the information is. Thus a respondent completing a questionnaire that bears absolutely no identifying marks – names, addresses, occupational details or coding symbols – is ensured complete and total anonymity. Non-traceability is an important matter, and this extends to aggregating data in some cases, so that an individual's response is unknowable or ensuring that data cannot be combined and individuals identified (Raffe *et al*., 1989) (e.g. in a school there may be only one middle-aged male teacher of religious education).

The principal way of ensuring anonymity, then, is removing any means of identification. Further strategies for achieving anonymity have been listed by Frankfort-Nachmias and Nachmias (1992), for example, the use of: (a) aliases and pseudonyms; (b) codes for identifying

people (to keep the information on individuals separate from access to them); and (c) password-protected files. Plummer (1983), likewise, refers to life studies in which names have been changed, places shifted and fictional events added to prevent acquaintances of participants discovering their identity. In experimental research the experimenter is interested in 'human' behaviour rather than in the behaviour of specific individuals (Aronson and Carlsmith, 1969, p. 33), and the researcher has no interest in linking the person as a unique, named individual to actual behaviour, so the research data can be transferred to coded, unnamed data sheets.

Hammersley and Traianou (2012) note that if anonymity and confidentiality cannot be guaranteed, then it should not be promised (p. 129). In an age of electronic data storage, this is a reality. Further, some participants may deliberately wish to be identified, and the researcher has to consider how to take account of this. It may be that the researcher accedes to the request or denies it (e.g. if the researcher wishes to avoid any risk of libel or considers that the purpose of the research is to produce knowledge and answer research questions, not to serve individuals' or groups' interests or to give them a voice) (p. 130).

7.10 Confidentiality

One way of protecting a participant's right to privacy is through the promise of confidentiality: not disclosing information from a participant in any way that might identify that individual or that might enable the individual to be traced. It can also mean not discussing an individual with anybody else or passing on the information to others in any form that can identify individuals. This means that although researchers know who has provided the information or are able to identify participants from the information given, they will in no way make the connection known publicly; the boundaries surrounding the shared secret will be protected. The essence of the matter is the extent to which investigators keep faith with those who have helped them.

It is generally at the access stage or at the point where researchers collect their data that they make their position clear to the hosts and/or participants. They will thus be quite explicit in explaining to participants what the meaning and limits of confidentiality are in relation to the particular research project. On the whole, the more sensitive, intimate or potentially discrediting the information, the greater is the obligation on the researcher's part to make sure that guarantees of confidentiality are carried out in spirit and letter. Promises must be kept.

Kimmel (1988) notes that some potential respondents in research on sensitive topics will refuse to cooperate when an assurance of confidentiality is weak, vague, not understood or thought likely to be breached. He concludes that the usefulness of data in sensitive research areas may be seriously affected by the researcher's inability to provide a credible promise of confidentiality. Assurances do not appear to affect cooperation rates in innocuous studies, perhaps because, as Kimmel suggests, there is an expectation on the part of most potential respondents that confidentiality will be protected.

A number of techniques have been developed to allow public access to data and information without confidentiality being betrayed. These have been listed by Frankfort-Nachmias and Nachmias (1992) as follows:

- deletion of identifiers (e.g. deleting the names, addresses or other means of identification from the data released on individuals);
- crude report categories (e.g. releasing the year of birth rather than the specific date, profession but not the speciality within that profession, general information rather than specific);
- microaggregation (i.e. the construction of 'average persons' from data on individuals and the release of these data, rather than data on individuals);
- error inoculation (deliberately introducing errors into individual records while leaving the aggregate data unchanged).

Cooper and Schindler (2001, p. 117) suggest that confidentiality can be protected by obtaining signed statements indicating non-disclosure of the research; restricting access to data which identify respondents, seeking the approval of the respondents before any disclosure about respondents takes place, non-disclosure of data (e.g. subsets that may be able to be combined to identify an individual).

Confidentiality also has to respect legal requirements. For example, if a child indicates that she is considering suicide or is at risk from an abusive parent, the researcher may have a legal obligation to inform relevant authorities. The researcher will need to make this clear, for example, at the start of an interview.

7.11 Against privacy, confidentiality and anonymity

Whilst a deontological and 'virtue ethics' approach to the ethics of educational research might demand that

rights to privacy be respected, on the other hand a utilitarian, consequentialist approach might argue that privacy could be violated if it is for the public good. Lincoln (1990) suggests that privacy protects the powerful and reproduces inequalities of power, whilst Howe and Moses (1999, p. 43) give examples where privacy should not be able to cloak wrongdoing (e.g. 'an abusive teacher … a sexist curriculum').

Wiles *et al*. (2008a, p. 419) indicate some of the complexities of ethical issues in confidentiality in their discussion of whether confidentiality should be broken in the interests of public or private safety, issues of actual or predicted criminal activity, if a person is at risk (e.g. a child who reports being abused, and legislation *requires* the reporting of this), and with vulnerable groups such as children, those with special needs, the recently bereaved, children whose parents have separated or who come from violent families. In many cases the researcher makes it clear before any interview commences that any information of a legal nature may be disclosed if the interviewer thinks the interviewee is at risk or if there is a legal matter at stake. However, it is not always as simple as this, as an interviewee may reveal some information that had not been anticipated. In other cases the researcher may want to give advice to a participant about seeking counselling or therapy. (Oliver (2003) cautions that the researcher is not herself/himself a counsellor or therapist (p. 71).)

In the case of covert research, there are no guarantees of confidentiality given in the first place. At issue here is where the duty of the researcher lies – to the research, to the individual, to the public or to whom – and the possible tensions between illegality, morality and the need to bring a matter to public awareness or knowledge. For example, if one is deliberately researching criminal activity it may be necessary to ensure confidentiality (maintaining 'guilty knowledge') or else the research will not take place at all. What does the researcher do if a court order is issued that requires the release of the data?

If researchers decide to opt for confidentiality then this can place them in a difficult situation where the research is emotionally draining, because, as Wiles *et al.* (2008a, p. 421) remark, it means that researchers cannot 'offload' their difficulties onto any other person.

Walford (2005, pp. 84–5) suggests that, whilst confidentiality, anonymity and non-traceability may be accepted or desirable norms for educational research, in some cases these norms may not apply or be achievable. For example, some participants or institutions may wish, or have a right, to be identified, as it might advance their cause or institution. Schools and headteachers might welcome publicity (Oliver, 2003, p. 77). As Wiles *et al.* (2008, p. 426) remark, in an age of increasing individualization, some individuals will insist on being identified. Further, the researcher is placed in a difficult situation with regard to confidentiality if a participant comments about another person who is not in the research and/or from whom no informed consent has been sought or obtained (Crow *et al*., 2006, p. 92): does the investigator use the data? Is it fair to exclude or include data about a third party because that third party has not been approached for informed consent?

Anonymity is also a double-edged sword. Whilst it might protect people, that may not be the main question; rather the question should be 'protect them from what?', as anonymity might become a cloak behind which participants can hide whilst making a range of negative, unsupported or even slanderous or libellous comments (cf. Oliver, 2003, p. 81). Maybe confidentiality and anonymity are only confined to certain forms of research.

More problematic is the question of what confidentiality actually means if the data are to be used for the research; if data are to be confidential and cannot be used or passed on, then what is the point of collecting or having the data? In this case it is perhaps anonymity and non-traceability that should be addressed rather than confidentiality, or the *scope* of confidentiality (its boundaries) should be clarified rather than a guarantee be given of absolute confidentiality (e.g. Oliver, 2003, p. 15).

It is often simply impossible to guarantee the anonymity of a person or an institution, as people can reassemble or combine data to identify a person or an institution or an institution can be identified by the 'locals', or indeed it can be identified by entering a few simple keywords from the research into an Internet search (Walford, 2005). Oliver (2003, p. 80) writes that it is impossible to give absolute guarantees of anonymity, especially where certain individuals are in named posts (e.g. a school principal). Here the commonly used advocacy of pseudonyms is no guarantee of anonymity. Walford (2005, p. 88) argues that promises of anonymity are often used by the researcher in order to gain access, though anonymity cannot actually be guaranteed, and, hence, it is ethically questionable whether anonymity should be promised. Whilst anonymity may bring data that are richer, keener and more acute or poignant than more anodyne research data which are given without promises of anonymity, this is not necessarily a justification for making promises of anonymity that cannot be kept.

Many devices can be used in the protection of anonymity, to 'put people off the scent', for example,

using pseudonyms, reporting a different geographical location from the one in which the research is actually carried out, providing misinformation (deliberately giving incorrect details of ages or sex), concealing identifying details (cf. Howe and Moses, 1999, p. 45), i.e. moving from 'disguise' to 'distortion' (Wiles *et al*., 2008, p. 422), in short removing context and, further, not always indicating that this has been done. However, this is problematic, as not only does it smack of telling lies and dishonesty, but it actually removes some of the very contextual data that are important for the research (Walford, 2005, p. 90), particularly for ethnographic research. To omit such necessary contextual details for researchers to understand the situation gives a spurious generalizability to the research. Walford suggests that it may be important to identify institutions and individuals, but that they should be given the right to reply in the research report, though this in turn is problematic. Who has the right to reply (all the participants?); what if very different replies are given? How are transparency, frankness and trust addressed in the relations between researchers and participants? These are knotty problems.

Howe and Moses (1999, pp. 44–5) make a cogent case against privacy and confidentiality, arguing that the 'thick descriptions' of interpretive research require a level of detail that cannot be obtained if privacy, confidentiality and anonymity are required. They argue that, as descriptions move towards becoming more 'objective', they become more anodyne and lose the very richness that they are intended to demonstrate, i.e. they become 'thin'.

7.12 Deception

The use of deception in social research has attracted considerable publicity and different opinions on its acceptability (British Educational Research Association, 2011; American Educational Research Association, 2011; Economic and Social Research Council, 2015). Is it acceptable to deceive people – by commission or omission – and, if so, under what circumstances, or is deception simply 'off the agenda'?

Concealing information

Deception resides in not telling people that they are being researched (in some people's eyes this is tantamount to spying), not telling the truth, withholding some or all information about the research, telling lies, 'giving a false impression' and 'failing to correct misconceptions' (Hammersley and Traianou, 2012c, p. 97), compromising the truth or withholding opinions.

Deception is applied to that kind of experimental situation where the researcher knowingly conceals the true purpose and conditions of the research, or else positively misinforms the participants, or exposes them to unduly painful, stressful or embarrassing experiences, without the participants having knowledge of what is going on. The deception lies in not telling the whole truth. Deception is a matter of degree, and is for the researcher to judge.

Advocates of the method feel that if deception is the only way to discover something of real importance, the truth so discovered is worth the lies told in the process so long as no harm comes to the participants (see the codes of ethics introduced earlier). Deception may be justified on the grounds that the research serves the public good, that the deception prevents any bias from entering the research and that it protects the confidentiality of a third party (e.g. a sponsor). The problem from the researcher's point of view is: what is the proper balance between the interests of science and the thoughtful, humane treatment of people who, innocently, provide the data?

The pervasiveness of the issue of deception becomes even more apparent when we remember that it is even built into many measurement devices, since it is important to keep the respondent ignorant of the personality and attitude dimensions that we wish to investigate. There are many problems that cannot be investigated without deception and, although there is some evidence that most participants accept without resentment the fact of having been duped once they understand the necessity for it (e.g. the Milgram obedience-to-authority experiment, see Chapter 30), it is important to keep in the forefront of one's mind the question of whether the amount and type of deception is justified by the significance of the study and the unavailability of alternative procedures.

Kelman (1967) has suggested three ways of dealing with the problem of deception. First, it is important that we increase our active awareness that it exists as a problem. It is crucial that we always ask ourselves the question of whether deception is necessary and justified. We must be wary of the tendency to dismiss the question as irrelevant and to accept deception as a matter of course. Active awareness is thus in itself part of the solution, for it makes the use of deception a focus for discussion, deliberation, investigation and choice.

The second way of approaching the problem concerns counteracting and minimizing the negative effects of deception. For example, participants must be selected in a way that will exclude individuals who are especially vulnerable; any potentially harmful manipulation must be kept to an acceptable level of intensity;

researchers must be sensitive to danger signals in the reactions of participants and be prepared to deal with crises as soon as, or before, they arise; and at the conclusion of the research, they must take time not only to reassure participants, but also help them work through their feelings about the experience to whatever degree may be required (see the discussions of the Milgram experiments and Stanford Prison Experiment in Chapter 30). The principle that participants ought not to leave the research situation with greater anxiety or lower levels of self-esteem than they came with is a useful one (the issue of non-maleficence again). Desirably, participants should be enriched by the experience and should leave it with the feeling that they have learned something.

The third way of counteracting negative effects of research employing deception is to ensure that adequate feedback is provided at the end of the research or research session. Feedback must be kept inviolable and in no circumstances should participants be given false feedback or be misled into thinking they are receiving feedback when the researcher is in fact introducing another experimental manipulation. Debriefing (see also Chapter 30) may include (Cooper and Schindler, 2001, p. 116):

- explaining any deception and the reasons for it;
- description of the purposes, hypotheses, objectives and methods of the research;
- sharing of the results after the research;
- follow-up psychological or medical attention after the research.

Even here, however, there are dangers. As Aronson and Carlsmith (1969) indicate, debriefing a participant by exposing her/him to the truth can be harmful than no debriefing; there is 'nothing magically curative about the truth' (p. 31), and great care has to be taken to ensure that the participant does not leave more uncomfortable than at the start of the experiment. They consider that the one essential aspect of the debriefing process is that researchers communicate their own sincerity as scientists seeking the truth and their own discomfort about the fact that they found it necessary to resort to deception in order to uncover the truth. As they say, 'no amount of postexperimental gentleness is as effective in relieving a subject's discomfort as an honest accounting of the experimenter's *own* discomfort in the situation' (Aronson and Carlsmith, 1969, pp. 31–2).

Another way of dealing with the problem of deception is to ensure that new procedures and novel techniques are developed so that deception becomes unnecessary. It is a question of tapping one's own creativity in the quest for alternative methods. It has been suggested that role-playing, or 'as-if' experiments, could prove a worthwhile avenue to explore (see Chapter 30). Here the participant is asked to behave as if he/she were a particular person in a particular situation. Whatever form they take, however, new approaches will involve a different set of assumptions about the role of the participant in this type of research. They require us to *use* participants' motivations rather than bypassing them.

Kimmel (1988) claims that few researchers feel that they can do without deception entirely, since the adoption of an overtly conservative approach could render the study of important research hardly worth the effort. A study of prejudice, for example, accurately labelled as such, could affect the behaviour of the participants. Deception studies, he considers, differ so greatly that even the harshest critics would be hard pressed to state unequivocally that all deception has potentially harmful effects on participants or is wrong. Indeed whilst research associations may discourage deception in research, they also recognize that, in some cases, it may be necessary and useful – the only way possible – but this requires careful justification.

Covert research

In the social sciences, the dilemma of deception has played an important part in research where participants are not told the true nature of the research, or where researchers conceal their identities and 'con' their way into groups, for example, alien, marginal, stigmatized or oppositional groups: the overt/covert debate (Mitchell, 1993). Covert or secret participation refers to that kind of research where researchers spend an extended period of time in particular research settings, concealing the fact that they are researchers and pretending to play some other role.

Bulmer (1982) notes that there are no simple and universally agreed answers to the ethical issues that covert research produces. Hornsby-Smith (1993, p. 65) argues that covert research violates informed consent, invades personal privacy, deceives people, risks harming participants when the research is published (e.g. Scheper-Hughes, 1979) and impairs the likelihood of other researchers researching the issue in the future or, indeed, of being able to conduct research, not least when overt research might have been used instead.

Douglas (1976a), Bulmer (1982) and Mitchell (1993), by contrast, argue that covert observation is necessary, useful and revealing, and that the most compelling argument in favour of covert research is that it has produced high-quality social science and has

advanced our understanding of society, which would not have been possible without the method. Indeed that is the view often taken in published codes of ethics (discussed earlier).

Covert research may be justified, for example, if the important data gathered could not have been gathered in any other way, or if it is necessary in order to gain access to organizations which would deny access (Mitchell, 1993; British Educational Research Association, 2011; American Educational Research Association, 2011), or to uncover questionable practices that, otherwise, would not come to light (e.g. child abuse) (cf. Oliver, 2003, p. 6). The consequentialist (e.g. utilitarian) argument for covert research is powerful.

7.13 Gaining access and acceptance into the research setting

The relevance of the principle of informed consent becomes apparent at an early stage of the research project – that of access to the institution or organization where the research is to be conducted and acceptance by those whose permission is needed before embarking on the task. Early access and acceptance offers the best opportunity for researchers to present their credentials as serious investigators and establish their own ethical position with respect to their proposed research.

Investigators cannot expect access as a matter of right. They have to demonstrate that they are worthy, as researchers and human beings, of being accorded the facilities needed to carry out their investigations. The advice of Bell (1991, p. 37) is to gain permission early on, with fully informed consent, indicating to participants the possible benefits of the research.

The first stage involves the gaining of official permission to undertake one's research in the target community. This will mean contacting, in person or in writing, an appropriate official, for example, the headteacher/principal. At a later point, significant figures who will be responsible for, or assist in, the organization and administration of the research will also need to be contacted – the deputy head or senior teacher, for instance, and certainly the class teacher if children are to be involved in the research. Since the researcher's potential for intrusion and perhaps disruption is considerable, amicable relations should be fostered as expeditiously as possible. If the investigation involves teachers as participants, propositions may have to be put to the stakeholders and conditions negotiated. Where the research is to take place in another kind of institution, the approach will be similar, although the organizational structure will be different.

Achieving goodwill and cooperation is especially important where the proposed research extends over a period of time: days, perhaps months in the case of an ethnographic study, or perhaps years where longitudinal research is involved. Access does not present quite such a problem when, for example, a one-off survey requires respondents to give up half-an-hour of their time or when a researcher is normally a member of the organization in which the research is taking place (an insider), though in the case of the latter, it may be unwise to take cooperation for granted. Where research procedures are extensive and complicated, however, or where the design is developmental or longitudinal, or where researchers are not normally based in the target community, problems of access are more involved and require greater preparation.

Having identified the official and significant figures whose permission must be sought, and before actually meeting them, researchers will need to clarify in their own minds the precise nature and scope of their research. They should have a total picture of what it all entails, even if the overall scheme is a provisional one (though we have to bear in mind that this may cause difficulties later). In this respect researchers could, for instance, identify: the aims of the research and its practical applications, if any; the design, methods and procedures to be used; the nature and size of samples or groups; what tests are to be administered and how; what activities are to be observed; which participants are to be interviewed; observational needs; the time involved; the degree of disruption and intervention envisaged; arrangements to guarantee confidentiality with respect to data (where necessary); the role of feedback and how findings can best be disseminated; the overall timetable within which the research is to be encompassed; and whether assistance will be required in the organization and administration of the research.

By such planning and foresight, both researchers and institutions will have a good idea of the demands likely to be made on both participants and organizations. It is also a good opportunity to anticipate and resolve likely problems, for example, those of a practical kind. A long, complicated questionnaire, for example, may place undue demands on the comprehension skills and attention spans of a particular class of nine-year-olds, or a relatively inexperienced teacher could feel threatened by sustained research scrutiny. Once this kind of issue has been resolved, researchers will be in a stronger position to discuss their proposed plans in an informed, open and frank manner (though not necessarily too open, see below) and may thereby more readily gain permission, acceptance and support. It must be remembered that hosts will have perceptions

of researchers and their intentions and that these need to be positive. Researchers can best influence such perceptions by presenting themselves as competent, trustworthy and accommodating.

Once this preliminary information has been collected, researchers are duly prepared for the next stage: making actual contact in person, perhaps after an introductory letter, telephone call or email, with appropriate people in the organization with a view to negotiating access. If the research is university-based, they will have the support of their university (and, where relevant, their supervisor). Festinger and Katz (1966) consider that there is real economy in going to the very top of the organization or system in question to obtain assent and cooperation. This is particularly so where the structure is clearly hierarchical and where lower levels are always dependent on their superiors. They consider it likely that the nature of the research will be referred to the top of the organization sooner or later, and that there is a much better chance of a favourable decision if leaders are consulted at the outset. It may also be the case that heads will be more open-minded than those lower down, who, because of insecurity, may be less cooperative.

The authors also warn against using the easiest entrances into the organization when seeking permission; researchers may perhaps seek to come in as allies of individuals or groups who have a special interest to exploit and who see research as a means to their ends, rather than entering the situation in the common interests of all parties, with findings equally available to all groups and persons. Investigators should seek as broad a basis for their support as possible. Other potential problems may be circumvented by making use of accepted channels of communication in the institution or organization. Festinger and Katz (1966) caution that if information is limited to a single channel then the study risks becoming identified with the interests that are associated with that channel.

Following contact, there is likely to be a negotiation process. At this point researchers will give as much information about the aims, nature and procedures of the research as is appropriate. This is very important: information that may prejudice the results of the investigation may have to be withheld. Aronson and Carlsmith (1969), for instance, note that one cannot imagine researchers who are studying the effects of group pressure on conformity announcing their intentions in advance. On the other hand, researchers may find themselves on dangerous ground if they go to the extreme of maintaining a 'conspiracy of silence', because, as Festinger and Katz (1966) note, such a stance is hard to keep up if the research is extensive and lasts over several days or weeks, and trying to preserve secrecy might lead to an increase in the spread and wildness of rumours. If researchers do not want their potential hosts and/or participants to know too much about specific hypotheses and objectives, then a way forward is to present an explicit statement at a fairly general level with one or two examples of items that may not be crucial to the study as a whole, though whether this constitutes deception is, itself, an ethical dilemma.

As most research entails some risks, especially where field studies are concerned, and as the presence of an observer scrutinizing various aspects of community or school life may not be relished by all in the group, investigators must at all times manifest a sensitive appreciation of their hosts' and participants' position and reassure anyone who feels threatened by the work. Such reassurance could take the form of a statement of conditions and guarantees given by researchers at this negotiation stage. By way of illustration, Box 7.4 contains conditions laid down for the Open University students' school-based research project.

At the stage of access and acceptance, situated ethics will determine what is acceptable and what is not acceptable.

BOX 7.4 CONDITIONS AND GUARANTEES PROFFERED FOR A SCHOOL-BASED RESEARCH PROJECT

1. All participants must be given the chance to remain anonymous.
2. All data must be given strict confidentiality.
3. Interviewees should have the chance to verify statements at the stage of drafting the report (respondent validation).
4. Participants should be given a copy of the final report.
5. Permission for publication must be gained from the participants.
6. If possible, the research report should be of benefit to the school and participants.

Source: Adapted from Bell (1991)

BOX 7.5 NEGOTIATING ACCESS CHECKLIST

1. Clear official channels by formally requesting permission to carry out your investigation as soon as you have an agreed project outline.
2. Speak to the people who will be asked to cooperate.
3. Submit the project outline to the head, if you are carrying out a study in your or another educational institution.
4. Decide what you mean by anonymity and confidentiality.
5. Decide whether participants will receive a copy of the report and/or see drafts or interview transcripts. There are cost and time implications. Think carefully before you make promises.
6. Inform participants what is to be done with the information they provide.
7. Prepare an outline of intentions and conditions under which the study will be carried out to hand to the participants.
8. Be honest about the purpose of the study and about the conditions of the research. If you say an interview will last ten minutes, you will break faith if it lasts an hour. If you are conducting the investigation as part of a degree or diploma course, say so.
9. Remember that people who agree to help are doing you a favour. Letters of thanks should be sent, no matter how busy you are.
10. Never assume 'it will be all right'. Negotiating access is an important stage in your investigation. If you are an inside researcher, you will have to live with your mistakes, so take care.

Source: Adapted from Bell (1991)

A pilot study can be useful to judge the effects of a piece of research on participants (Oliver, 2003, p. 37). Where a pilot study is not feasible it may be possible to arrange one or two scouting forays to assess possible problems and risks. By way of summary, we refer the reader to Box 7.5.

Access may not be a once-and-for-all matter. For instance, in longitudinal studies, say of a school, access may become a problem if the researcher encounters new students, new parents and new staff, and access may have to be renegotiated (cf. Brooks *et al*., 2014, p. 157).

7.14 Power and position

The researcher is often seen to be, or is, in an asymmetric position of power with regard to the participants; the former may have more power than the latter, be this by status, position, knowledge, role or whatever. The researcher typically determines the agenda, the timing and duration of the research and, for example, interviews, what counts as acceptable and useful data, to whom the data are released, who might or might not be identifiable, and so on. As Brooks *et al*. (2014) remark, 'power relations are immanent in all research settings' (p. 106), and researchers may occupy different social and power positions from participants.

This is particularly the case when researching with children, as they are more vulnerable and, in many settings, more powerless than adults or researchers. For example, Morrison (2013a) reports on interviewing children in a situation in which: there were strong asymmetries of power and age; the agenda was decided by evaluators-as-interviewers; and semi-structured qualitative interviews operated in a strongly focused and question-and-answer style (pp. 320–1). He reports several strategies used to overcome the strangeness of the situation and the power differentials, to put students at ease and treat them as important, indeed to make the interviews 'a positive and enjoyable experience for the children', so that they would leave the interviews 'feeling positive about themselves and the interviews' (p. 321) (see also Chapter 25 on interviewing children).

One typical response to asymmetries of power is to try to reduce the power differentials, enabling participants to have power over decision making in the research. However, researchers have to consider the limits of this; Hammersley and Traianou (2012), for example, ask whether rapists and paedophiles should be accorded equal powers to the researcher (p. 82). Another is to establish rapport and trust, which might take the form of ensuring a match between the characteristics of the researcher and the participants (e.g. age, gender, ethnicity, language, background, biography etc.). This might be particularly important in researching minority or marginalized, excluded groups, i.e. those with limited perceived agency or power (Brooks *et al*., 2014, pp. 112–13).

Hochschild (2012) notes that 'emotion work' in research involves dealing with the emotions of others and, as part of this, requires researchers to keep their own true emotions in check, to some extent setting aside their own emotions in handling those of participants. They must be emotionally detached yet friendly and positive, and in the research situation, particularly, for example, in-depth interviews about sensitive matters, this requires an ability to be empathetic and suitably informal and yet formal. Hammersley and Traianou (2012) note that researchers may have to be prepared to tolerate behaviour, attitudes and opinions that they personally detest or find unacceptable (p. 55) in order to conduct valuable and valid research.

Researchers, then, have to be acutely aware of possible or likely asymmetries of power and take appropriate steps to address the ethical issues that such awareness raises.

7.15 Reciprocity

Reciprocity means giving or giving back something to the participants in the research in return for their participation. Researchers should never lose sight of the obligations they owe to those who are helping; an ethical matter.

Sikes (2006, p. 112) quotes the words of Lather (1986) in describing 'rape research' as 'research in which the researcher gets what they want and then clears off, giving little or nothing in return and maybe even causing damage' (see also Reinharz, 1979). This is unethical. Similarly, Laing (1967, p. 53) reminds us that 'data' are 'things that are given' – gifts – rather than 'things that are captured' (i.e. 'data' rather than '*capta*'). The researcher has some obligation to give something back to the participants.

A researcher may gain promotion, publications, a degree, research sponsorship and celebrity from a piece of research. However, the research might still leave the participants untouched, underprivileged, living and working in squalid and under-resourced conditions, under-supported and with no material, educational or other improvements brought to the quality of their lives and work. As one of Whyte's contacts remarked ruefully in his celebrated study of an Italian slum in *Street Corner Society* (1955), the locals had helped many researchers to become famous and get their doctorates, though leaving the locals' quality of life with no improvement (see also Willis and Saunders (2007, p. 96), reporting on indigenous populations who had been incessantly and minutely interrogated by outside 'experts' and left impoverished).

Baumrind (1964) warns of the possible failure on the researchers' part to perceive a positive indebtedness to their participants for their services, perhaps because the detachment which investigators bring to their task prevents appreciation of participants as individuals. This kind of omission can be averted if the researchers are prepared to spend a few minutes with participants afterwards in order to thank them for their participation, answer their questions, reassure them that they did well and generally talk to them for a time.

The issue is also raised here of whether participants should be given inducements to participate, for example, payment, gifts or the opportunity to enter a 'lucky draw'. A different kind of inducement to participate may be in the form of advice to participants or, for example, educational advice to parents. On the one hand, the argument runs that any kind of material inducement distorts a genuine relationship between the researcher and the participants, such that participants may say something or join the research because they will be paid for it, or may give perfunctory information just to be able to obtain the reward, and whose commitment is actually very small. On the other hand, participants are giving their time and effort to the research, so they should be paid for it, just as in other kinds of work (cf. Oliver, 2003, pp. 23, 59). Head (2009) notes that paying participants is widespread in medical and psychological research, indeed is ethically desirable in equalizing (power) relationships between researcher and participants, and it can apply to qualitative research as well, as it encourages participation and response rates. Payment should be commensurate with the amount of time and effort expended, and should not be coercive or corrupting (pp. 340–3).

Brooks *et al*. (2014) suggest that it may be acceptable to offer some incentives, but not to the extent that this is likely to distort the research or to have participants who join the research for the sake of receiving the incentive on offer (p. 97). On the other hand, they note that inducements may discourage participation, depending on what those inducements might be, for example, some parents may not wish to have meal vouchers as they would be seen as being in need of such vouchers and, hence, embarrassed (pp. 98–9). They also raise the question of to whom to offer the incentives, for example, the child, the parents, the school (p. 97).

7.16 Ethics in data analysis

Data analysis must be ethical. It must not mis-present findings or the phenomenon itself, and such misrepresentation can happen in many ways, for example:

- using inappropriate data-analysis techniques;
- being unfairly selective with regard to the data used;
- falsifying and making up data;
- ignoring, omitting or concealing data that do not 'fit' what the researcher wishes to show, or using data to support a preconceived or preferred view;
- being unfair to the data and what they show, for example, misrepresenting what the data are saying or showing;
- overstating and understating what the data show, and over-interpreting data and pieces of data;
- giving undue weight and priority to some data;
- projecting one's own values onto the data, and presenting the researcher's own views and own preferred frameworks for data analysis which distort the analysis;
- using inappropriate statistics, or collapsing, over-reducing and over-summarizing data;
- selecting statistics which show the situation in a better or worse light than is really the case;
- ignoring outliers;
- making the false claim that large samples prove reliability and validity;
- making false claims of causality;
- failing to exert suitable controls in the data analysis;
- breaching the ethical requirements of confidentiality and anonymity (e.g. in visual data);
- editing out items in visual data (cropping and recolouring);
- failing to give sufficient 'voice' to participants, for example, in qualitative research;
- making false, exaggerated, sensationalized, scandalized and unsubstantiated claims from the data;
- failing to consider rival interpretations and explanations of the findings;
- judging rather than analysing the data.

Whilst it is almost impossible for researchers to free themselves from their values and perspective in a post-positivist era, and indeed there may be unintentional breaches of ethics, researchers must be vigilant, very self-aware and reflexive in their data analysis. It is not true, for example, that statistics are self-justifying: the researcher has immense control over which statistics to use and what they might or might not show. Further, in mixed methods research different sample sizes may be used, and care has to be exercised not to focus too heavily on large samples to the detriment of small samples (Creswell, 2012, p. 553).

Ethics also features in discussions of ownership of the data, for example, when the ownership passes from the participants to the researcher and with what constraints, requirements, conditions and powers over the use and dissemination of the findings required by the participants (cf. Howe and Moses, 1999, p. 43; Brooks *et al*., 2014). Researchers need to be clear whether they own the data once the data have been given, or whether the participants have control over what is released and to whom; this should be agreed, where possible, before the research commences. Oliver (2003, p. 63), for example, argues that the raw data are still the property of the participants, but once the data have been analysed and interpreted, they become the property of the researcher. This is unclear, however, as it does not cover, for example, observational data, field notes and the like, which are written by the researcher, though often about other people. Negotiating ownership rights, rights to release or withdraw data, rights to control access to data, rights to verify and validate data, rights to vet data or see interim or incomplete or uncompleted reports, rights to select data and decide on their representativeness, rights to own or change the final report and rights to retain data after the research (e.g. for other purposes, as in the ongoing compilation of a longitudinal or comparative study) moves the conduct of research beyond being a mechanical exercise to being an ethical exercise (cf. Oliver, 2003, pp. 63–5).

Disclosure and data usage

The researcher will frequently find that disclosure impinges on methodological and ethical issues (Hitchcock and Hughes, 1989). They pose questions that may arise in such a situation. 'Where, for the researcher, does formal observation end and informal observation begin?' 'Is it justifiable to be open with some teachers and closed with others?' 'How much can the researcher tell the students about a particular piece of research?' 'When is a casual conversation part of the research data and when is it not?' 'Is gossip legitimate data and can the researcher ethically use material that has been passed on in confidence?' The list of questions is endless yet they can be related to the nature of both the research technique involved and the social organization of the setting being investigated. One key to the successful resolution of such questions may lie in establishing good relations, involving the development of a sense of rapport between researchers and participants that leads to feelings of trust and confidence.

Finch (1985, pp. 116–17) comments on the possibly acute political and ethical dilemmas arising from how data are used, both by the researcher and others, and the researcher has a duty of trust placed in him/her by the participants to use privileged data appropriately, not least for improvement of the condition of the participants.

BOX 7.6 ETHICAL PRINCIPLES FOR THE GUIDANCE OF ACTION RESEARCHERS

Observe protocol: Ensure that the relevant persons, committees and authorities have been consulted, informed and that the necessary approvals/permissions have been obtained.
Involve participants: Encourage potential stakeholders in the improvement to be involved in the project.
Negotiate with those affected: Take account of the responsibilities and wishes of participants, as not all of them may wish to be involved directly.

Report progress: Keep the work visible and be open to suggestions, to take account of unforeseen and unseen implications or outcomes; enable colleagues to have the opportunity to challenge or lodge a protest.
Obtain explicit authorizations: For example, if you wish to observe your colleagues and/or examine documents.

Negotiate descriptions of people's work: Always enable those described or identified in the research to challenge your accounts, for example, on grounds of fairness, relevance and accuracy.
Negotiate accounts of others' points of view (e.g. in accounts of communication): Enable participants in interviews, meetings and written exchanges to require amendments which improve fairness, relevance and accuracy.

Obtain explicit authorization before using quotations: For example, in using verbatim transcripts, attributed observations, excerpts of recordings (audio and video), judgements, conclusions or recommendations in reports.

Negotiate reports for various levels of release: Different audiences require different kinds of reports; what may suit an informal verbal report to a faculty meeting may not be suit a staff meeting, report to council, an academic article, a newspaper, a newsletter to parents; be conservative if it is not possible to control distribution.

Accept responsibility for maintaining confidentiality.
Retain the right to report your work: Provided that participants in the research are satisfied with the fairness, accuracy and relevance of accounts which pertain to them, and that these accounts do not unnecessarily expose or embarrass them, the accounts should not be subject to veto or sheltered by claims of confidentiality.

Make your principles of procedure binding and known: All those involved in the action research project must agree to the principles before the project commences; others must be aware of their rights in the project.

Source: Adapted from Kemmis and McTaggart (1981)

Box 7.6 presents a set of ethical principles specially formulated for action researchers by Kemmis and McTaggart (1981) and quoted by Hopkins (1985).

7.17 Ethics in reporting and dissemination

As with data analysis, the researcher has an ethical duty to ensure that the results of the research are reported fairly, credibly and accurately, without misrepresentation, unfair selectivity (exclusion and inclusion, or inappropriate piecemeal reporting of different parts, e.g. in different journals) (Creswell, 2012, p. 279), plagiarism, untenable claims, exaggeration or understatement, misinterpretation, bias and under-reporting or over-reporting certain findings to the detriment of a more balanced and fair view. The reporting must be honest, true, fair and in a format that the audiences of the research will be able to access and understand (e.g. lay or professional audiences). Further, potential conflicts of interest must be disclosed (many ethics committees require this).

Attention must also be given to confidentiality, anonymity and non-traceability, and this might extend to obtaining informed consent for dissemination and disclosure, which, in turn, raises issues of what informed consent should include and for how long it applies. Will an external, internal or local audience be able to identify the participants and institutions in the research, particularly if it is possible to combine data, or should deliberate attempts be made to disguise individuals and institutions, even to the point of fabricating details in order to put audiences 'off the scent'? This is particularly an issue if the research reports negative findings concerning individuals, groups, institutions and communities, i.e. where the research might cause harm.

Here, as earlier in this chapter, the researcher faces again the issue of what is in the public interest versus what respects the participants' privacy (Pring, 2015). Whilst Pring (2015) suggests that there is a prima facie case for the public's right to know – which is why research is undertaken in the first place – and whilst this breaks down the secrecy that often surrounds institutions which, in fact, should be publicly accountable, he also notes that the truth can hurt. Whose interests does the dissemination of the research protect or threaten? How are beneficence and non-maleficence interpreted and addressed? Is it acceptable for some individuals to be harmed if the greater public good is being served (i.e. the deontological view versus the utilitarian view of ethics)? The Nuremberg Code, for example, expressly argues against harming individuals in the pursuit of societal benefit (Farrimond, 2013, p. 27). Should the researcher be judgemental, and, if so, how, in whose interests and at whose expense? Should the researcher inform participants in advance of what will be disseminated and offer them the right of veto, or are the data, once given, the property of the researcher? Such issues become challenging when, for example, negative findings and divided loyalties are at stake, or if the research findings and dissemination might operate against the interests of the individual, group or community in question. As Brooks *et al*. (2014, p. 140) note, it is the researcher who stands to gain the most from the research but this does not preclude a duty of care and respect for participants and communities, and negative findings often shout louder than positive findings. This extends not only to research reports but to the data themselves, for example, written and visual data (e.g. photographs, video material which identify people).

Morrison (2006) considers the case of a school that is under-performing, poorly managed or badly led. Does not the 'consumer', indeed the state, have a right or a duty respectively to know or address this, such action typically involving the exposure to the public of a school's shortcomings, and will this not damage individuals in the school, the principal and the teachers? What 'fiduciary trust' (Mitchell, 1993) not to harm individuals (the ethical issue of 'non-maleficence') does the researcher have to the school or to the public, and how can these two potentially contradictory demands be reconciled? Should the researcher expose the school's weaknesses, which almost certainly could damage individuals but which may be in the public interest, or, in the interests of *primum non nocere*, remain silent? The pursuit of truth and the pursuit of trust may run counter to each other (Kelly, 1985, p. 147); indeed Kelly herself writes that 'I do not think we have yet found a satisfactory way of resolving this dilemma'.

In reporting and disseminating the research, the researcher needs to consider, even anticipate, who the audiences will be and the likely or possible effects on them of the reporting and dissemination. The research is only one interpretation of the findings, and the researcher has to make this clear (being reflexive), i.e. other voices might speak differently about the research. Researchers have to be mindful that once the research is in the public domain, they have no control over how it will be used.

The participants' sensibilities need also to be taken into account when the researcher comes to write up and disseminate the research. It is unacceptable for researchers to show scant regard for participants' feelings at the report stage. A related issue concerns the formal recognition of those who have assisted in the investigation, if such be the case. This can be done in a foreword, introduction or footnote. This means that the authors must consider acknowledging and thanking all who helped in the research, identifying by name those whose contribution was significant, but not if such identification jeopardizes previously agreed confidentiality and anonymity.

Personal data are defined in law as those data which uniquely identify the individual providing them. When such information is publicized with names through the media, for example, privacy is seriously violated. The more people there are who can learn about the information, the more concern there must be about privacy. This extends to the archiving of data, which should consider the removal of details which can identify individuals and institutions, as Freedom of Information Acts might permit the public to access archived data held by individuals, institutions and associations.

The term 'betrayal' is often applied to those occasions where data disclosed in confidence are revealed publicly in such a way as to cause embarrassment, anxiety or suffering to the participant disclosing the information. It is a breach of trust, in contrast to confidentiality. As Plummer comments, 'there is something slightly awry when a sociologist can enter a group and a person's life for a lengthy period, learn their most closely guarded secrets, and then expose all in a critical light to the public' (Plummer, 1983, p. 146). How does one write an honest but critical report of teachers' attitudes if one hopes to continue to work with those involved, for example, in action research (Kelly, 1989)?

Finch (1985) raises ethical issues in the consequences of reporting. In her research she worried that her reporting

> could well mean that I was further reinforcing those assumptions deeply embedded in our culture and political life that working class women (especially the urban poor) are inadequate mothers and too incompetent to be able to organize facilities that most normal women could manage.
>
> (p. 117)

Indeed she uses the word 'betrayal' (p. 118) in her concern that she might be betraying the trust of the women with whom she had worked for three years, not least because they were in a far worse economic and personal state than she herself was.

Whilst some researchers may place an embargo on having their research made available to the public (e.g. for five years) in order to protect participants (cf. Sikes, 2006, p. 111), this calls into question the values, purposes and ethical justifiability of research that cannot be disseminated and hence 'cannot contribute to the cumulativeness of knowledge' (p. 111), the latter being a signal feature of research (Pring, 2015).

7.18 Responsibilities to sponsors, authors and the research community

The researcher has responsibilities, indeed in many situations, obligations, to different parties and to legal regulation (Ary *et al*., 2002, pp. 504–7): sponsors, participants, stakeholders, authors and the research community. The researcher has to consider responsibility to the sponsors, and, again, this may pose a dilemma between what is in the public interest versus what is in the private or institutional interest or the interest of the sponsor. Sponsors may wish to restrict, prevent or censor dissemination, or control who sees what, when and in what form, and this might challenge academic freedom and fidelity to the phenomenon being researched. The sponsor may not wish to be identified or, indeed, may deliberately seek to be identified. This is a matter that should be agreed before the research commences, in order to avoid challenges arising too late in the research.

In reporting, some authors will be concerned about the order of the authors' names in, for example, an article or book. Who is the first-named, principal author? Should a research supervisor's name be included simply because he or she requires this, even though he or she has made no substantive contribution to the research, or should the supervisor's name be included in an acknowledgement (cf. Brooks *et al*., 2014, p. 148)? What are the politics involved in placing authors' names in a particular order? What are the consequences for authors (e.g. with regard to career promotion)?

The researcher also has responsibilities to the research community, for example, not to jeopardize the reputation of the research community (e.g. the university) or spoil the opportunities for further research. A novice researcher working for a higher degree might approach a school directly, using a clumsy approach, with inadequate data-collection instruments and a poor research design, and then proceed to publicize the results as though they are valid and reliable. At the very least the novice should have sought and gained advice from the supervisor, modified the research as necessary, gained approval for the research, made suitably sensitive overtures to the school and agreed rights of disclosure. Not to do so puts the researcher's (or others') institution at risk of being denied further access, of damaging the reputation of the institution, and, if word spreads, of being publicly vilified and denied the opportunity for further research to be conducted. In this case the novice researcher has behaved unethically.

Further, if a negative research report is released, will schools retrench, preventing future research in schools from being undertaken? Negative research data, such as reported evidence on deliberate grade inflation by schools in order to preserve reputation (Morrison and Tang, 2002), may not endear researchers to schools.

The researcher has a responsibility to colleagues to:

- protect their safety (e.g. in conducting sensitive research or research in dangerous locations);
- protect their well-being;
- protect their reputation;
- enable further research to be conducted;
- expect them to behave ethically;
- ensure that they adhere to correct and agreed procedures;
- protect the anonymity and confidentiality of sponsors if so agreed.

However, these may conflict with the public's right to know. The researcher, too, is a member of a research community, and this brings ethical responsibilities.

7.19 Conclusion

In this chapter we have attempted to acquaint readers with some of the ethical difficulties they are likely to experience in the conduct of research. It is not possible to identify all potential ethical questions or adjudicate on what is correct researcher behaviour. We have demonstrated that ethical principles are open to contestation, differences of interpretation and conflicts

between them; the researcher has to consider how ethical principles inform the particular research or research situation. Ethical principles help to guide researchers and only rarely definitively prescribe or proscribe research matters. In other words, whilst the knowledge of ethical principles help in all aspects and stages of the research, nevertheless ethics are 'situated' and particular to a specific situation. It is for the researcher to decide how to address and apply ethical principles in coming to a decision on how to act in the

BOX 7.7 ETHICAL PRINCIPLES FOR EDUCATIONAL RESEARCH (TO BE AGREED *BEFORE* THE RESEARCH COMMENCES)

Responsibility to research

The researcher should be competent and aware of what is involved in conducting research.

The research must be conducted rigorously and with the correct procedures – avoid misuse of procedures at all stages.

Report procedures accurately and publicly (rigour).

Don't jeopardize future research(ers).

Report clearly and make data available for checking.

Tell the truth (do not tell lies or falsify data, avoid being unfairly selective, e.g. to support a case, do not misrepresent data).

Maintain the integrity and autonomy of the research, for example, avoid censorship of, or interference with, the research by sponsors/those who give permission for the research to be undertaken.

Responsibility to participants and audience(s)

Gain fully informed consent where appropriate (usually in writing), in order to respect self-determination and autonomy; provide information on all aspects of the research and its possible consequences.

Decide whether, and how, overt or covert research is required/justified.

Decide whether, and how, deception is required/justified; be honest or justify dishonesty.

Ensure non-maleficence (no harm, hurt or suffering to be caused to participants and those who might be affected by the research); be humane.

Ensure beneficence (the research will bring benefit to the participants or will contribute to the welfare of participants).

Ensure that participants do not leave the research worse off than when they started it.

Respect people's rights and dignity and interests, and be respectful – research participants are subjects, not objects to be exploited. Treat people as subjects, not objects.

Agree individual's rights to privacy.

Ensure participants have the right to withdraw at any time.

Inform participants about who will have access to the data/report, i.e. the audiences of the research, how public it will be, when it will become public and how it will be disseminated; negotiate levels of release (i.e. who will see which parts of the research).

Ensure anonymity/confidentiality/non-traceability; if these are not possible then tell participants in advance.

Indicate how anonymity will be addressed (e.g. by confidentiality, aggregation of data).

Inform participants how data will be collected and how files/questionnaires/audio data/video data/computer files will be stored during the research and destroyed after use.

Ensure sensitivity to people (e.g. age, ethnicity, gender, culture, religion, language, socio-economic status etc.).

Gain permission from all relevant parties (e.g. parents/guardians, school, principals etc.) for access.

Respect vulnerability (e.g. in interviewing children/those without power).

Agree respondent validation.

Agree ownership of the data (and when ownership passes from participants to researcher).

Allow time for review.

Avoid causing unnecessary offence. Thank the participants.

Ensure that participants and sponsors have the right to dissent/distance themselves from the research.

Demonstrate social responsibility and obligations.

Consider indemnification, liabilities and disclaimers.

Don't abuse your position/power as a researcher.

Don't use dangerous methods.

specific research in question. Such decisions almost inevitably involve compromises.

Although no code of practice can anticipate or resolve all problems, there is a sixfold advantage in fashioning a personal code of ethical practice. First, such a code establishes one as a member of the wider scientific community having a shared interest in its values and concerns. Second, a code of ethical practice makes researchers aware of their obligations to their participants and also to those problem areas where there is a general consensus about what is acceptable and what is not. In this sense it has clarificatory value. Third, when one's professional behaviour is guided by a principled code of ethics, it is possible to consider that there may be alternative ways of doing the same thing; ways that are more ethical or less unethical should one be confronted by a moral challenge. Fourth, a balanced code can be an important organizing factor in researchers' perceptions of the research situation, and as such may assist them in their need to anticipate and prepare. Fifth, a code of practice validated by their own sense of rightness will help researchers to develop an intuitive sensitivity that will be particularly helpful to them in dealing with the unknown and the unexpected, especially where methods such as ethnography and participant observation are concerned. And sixth, a code of practice will bring discipline to researchers' awareness. Here Box 7.7 raises considerations to be borne in mind in planning, conducting and reporting research.

Box 7.7 raises issues and suggestions, not solutions or decisions. These latter two have to be decided by each researcher in respect of the particular situation he or she faces. Ethics are 'situated'. For a summary of ethical principles for social research and other ethical issues explored in this chapter, we refer readers to the companion website.

Note

1 The word 'subjects' is ambiguous: contrasted with 'objects' it could be positive, according equal status and respect to the participants; on the other hand it could be negative in that participants are subjected to the wishes of the researchers ('subject' literally means 'thrown under' or 'thrown below').

Companion Website

The companion website to the book provides additional material and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Ethics in Internet research

CHAPTER 8

The rise of Internet-based research, online research and virtual worlds has created a new site in which interactions take place between individuals and communities from ethically plural cultures and backgrounds. This raises many ethically ambiguous and contested issues which we introduce here. Many issues that we raise in Chapter 7 apply to Internet research, and we advise readers to review that chapter.

The following pages will explore:

- what Internet research is
- key ethical issues in Internet research
- informed consent
- public and private matters
- confidentiality and anonymity
- ethical codes for Internet research

8.1 What is Internet research?

Internet research is defined by Buchanan and Zimmer (2012) as that which uses the Internet to collect data using an online tool, or comprises studies of Internet use and how people use it, or which uses online datasets, databases or other materials (p. 2). Similarly the Association of Internet Researchers (2012) indicates that Internet research is that which:

a utilizes the Internet to collect data or information, e.g., through online interviews, surveys, archiving, or automated means of data scraping;
b studies how people use and access the Internet, e.g., through collecting and observing activities or participating on social network sites, listservs, websites, blogs, games, virtual worlds, or other online environments or contexts;
c utilizes or engages in data processing, analysis, or storage of datasets, databanks, and/or repositories available via the Internet;
d studies software, code, and Internet technologies;
e examines the design or structures of systems, interfaces, pages, and elements;
f employs visual and textual analysis, semiotic analysis, content analysis, or other methods of analysis to study the web and/or Internet-facilitated images, writings, and media forms;
g studies large-scale production, use, and regulation of the Internet by governments, industries, corporations, and military forces.

(Association of Internet Researchers, 2012, p. 3)

The British Psychological Society (2013) defines Internet research as 'any research involving the remote acquisition of data from or about human participants using the Internet and its associated technologies' (p. 3), which addresses both reactive and unobtrusive research. To assume homogeneity in Internet-based research is to misrepresent its diversity (Madge and O'Conner, 2005; Orton-Johnson, 2010).

Orton-Johnson (2010) and Jones (2011) note that the Internet is a tool, a means, a medium, a locale (a place to acquire and keep data), an object for research and a distribution channel for research. It includes data-collection instruments, web pages, chat rooms, blogs, email, discussion boards, virtual worlds, forums, social networking sites and pages, and so on – the list expands exponentially over time. It can enable unobtrusive research (where people do not know that their data are being collected, e.g. 'big data'; Beneito-Montagut, 2017) and intrusive research (where people are canvassed for their participation and/or data).

Whilst the Internet is global and not bounded by countries and territories, it operates differently in different jurisdictions and is regulated by differing laws in different parts of the world. Internet usage in research has exposed fissures in traditional conceptions of public and private spaces, and these, in turn, raise ongoing and emergent ethical questions. Indeed ethics has to play 'catch-up' in terms of online research (Convery and Cox, 2012).

8.2 What are key ethical issues in Internet research?

Internet research covers three main types (Farrimond, 2013): passive (the researcher is non-participant, e.g. studying data and sites on the Internet), active (researcher

is a participant, e.g. in an online community) and online traditional forms (e.g. surveys). Eysenbach and Till (2001) set out many key areas for ethical consideration in studying Internet communities: intrusiveness, perceived privacy, vulnerability, potential harm, informed consent, confidentiality and intellectual property rights. However, this is only a starting point, as Internet researchers must address the many issues that we raise below.

What do conventional conceptions of privacy, confidentiality, anonymity, ownership of intellectual property, vulnerability, harm, authenticity and informed consent really mean in a borderless world in which people are traceable yet never seen face-to-face, their data are tracked, recorded, aggregated, combined, stored indefinitely and interrogated without their knowledge, and where their private, even intimate, thoughts, communications and pictures are open to the public?

Buchanan and Ess (2009), surveying over 700 US ethics review boards, note that they were primarily concerned with matters of privacy, informed consent, confidentiality, security of data, and recruitment procedures. However, the field is wider than this. The Association of Internet Researchers (2012), Farrimond (2013), Barnes *et al*. (2015), Busher and James (2015), James and Busher (2015), Kontopoulou and Fox (2015), Roberts and Allen (2015), Stevens *et al*. (2015) suggest that ethical issues in Internet research have a huge embrace, here presented in alphabetical order:

- agency and the 'other' in online research (Busher and James, 2015, p. 170);
- beneficence and benefits (and for whom);
- blurring of online and real worlds: demarcation matters for privacy;
- combining online and face-to-face aspects of data collection, and the relationship between online and offline situations for participants and the researcher (Busher and James, 2015; James and Busher, 2015);
- conflicts of interest (where the researcher is a participant in an Internet group);
- consideration of online research by ethics committees;
- deontological and utilitarian issues;
- disclosure, data quality (e.g. representativeness of the sample) and veracity;
- ethical appraisal and approvals;
- the ethics of 'forced responses' (e.g. when a participant cannot proceed in a survey until all the questions on a screen have been answered);
- fairness;
- identity construction and protection, self-representation, authenticity, credibility and authentication;
- inapplicability of some traditional ethical guidelines and the rise of emergent challenges;
- informed consent, permissions and ensuring that participants know what they are consenting to, and the age of consent;
- keeping promises (fidelity);
- micro-celebrity status (Ramírez and Palu-ay, 2015, p. 146);
- opportunity for one participant to send in multiple completed online surveys;
- ownership of data and copyright concerns;
- power distribution, asymmetries of power and the need for justice;
- privacy, confidentiality and anonymity (e.g. the tracking of individuals online);
- private and public domains;
- privatized, deprivatized, semi-privatized, public, semi-public domains;
- questioning what counts as evidence;
- rapport in online research;
- reflexivity and transparency;
- representation;
- respondent validation;
- risk management, duty of care and protecting participants from harm and malicious intent;
- security;
- tensions between, and ambiguity in, private and public spheres;
- transparency;
- use of incentives;
- use of quotations that might be able to identify individuals through an internet search;
- uses of social media for research;
- virtual public spaces;
- visual data and their use.

These issues are addressed in our discussions that follow.

A twin guiding principle in Internet research, as with conventional research, is the avoidance of harm to people (non-maleficence) and the promotion of beneficence. The researcher must operate in what he or she considers are the best interests of participants. In this respect the rights of participants trump the rights or threats to the integrity of the research.

8.3 Informed consent

Informed consent is not straightforward in online research. For example, the researcher may not know who the actual person is who is answering, say, an online survey, and whether the details that they enter are honest and correct. How can informed consent be

gained from someone who is unseen and when there are no checks on whether the participant has understood the implications? Does one need consent from minors or their parents in online research? Can participants subsequently have their data withdrawn if they wish to withdraw from the research (Brooks *et al*., 2014, p. 93)? What if a person does not complete an online survey or withdraws from an ongoing piece of research: does the informed consent cease? And, anyway, how can the researcher trace which participants have given which data online?

Marshall and Rossman (2016, p. 183) note that researchers considering informed consent in Internet research face issues such as: whether, how much and in what sense and domains the data are public or private (who constitutes the research community); how sensitive the topics are; how much interaction will be required; the vulnerability of participants; and whether consent is actually necessary.

Seeking informed consent might come as an intrusive shock or a disruption to some participants, who had not realized that their data (e.g. from chat rooms, forums, social networking sites) were being monitored or collected. On the other hand, as with non-electronic research, covert research and deception might be justified in certain circumstances (Glaser *et al*., 2002), for example, where it is essential not to have informed consent for fear of 'blowing one's cover'. Indeed Denscombe (2014), reporting on Glaser's *et al*.'s (2002) study, notes that 'the respondents' statements were made in a public forum.… [T]he deception was absolutely necessary … and respondents' identities were carefully protected' (p. 322).

How easy, possible or realistic is it to obtain informed consent, and, if so, from whom (participants, parents, gatekeepers etc.)? How do we know that informed consent has really been given (Buchanan and Zimmer, 2012), when, for how long (e.g. in archived data), for what (use, and release, of data) and on whose behalf? It might be assumed that participants who complete an online survey, for example, are thereby giving consent, but have they really been informed about what they are consenting to and what might happen with the data? It may be that informed consent for Internet-based research has to be negotiated and renegotiated with participants over time as the research unfolds, and this may put off some participants.

Further, a researcher may not be able to identify a participant, such as, for example, in a survey conducted with non-disclosure of identifying details by the participant. Here informed consent is not possible, raising the question of whether the researcher uses or does not use the data (e.g. from chat rooms, forums, blogs, social networking sites)?

Ensuring informed consent may be obtained by requiring the researcher to provide information and asking participants to check an 'I accept' box, but this is akin to asking people to read the fine print of all the software that they download before checking the 'I accept' box, which typically they don't read. Some online research might ask participants to complete an 'I accept' box in a step-by-step staged process, whereby they are given some information on one sub-element or screen, to which they agree, and then later given information about the next sub-element or screen, to which they agree, and so on; this prevents the participant from being overwhelmed with too much information at once, but it may risk the participants dropping out if they are frequently having to check an 'I accept' box. Care must be taken to avoid long statements of information before checking a consent box, as participants may not read them. Some online research will place the consent box at the very end of, for example, the survey, so that participants know that they can draw back from sending data.

8.4 Public and private matters

Online ethical issues also arise in the context of 'big data'. With the rise of big data and online networking, data collection, storage and retrieval, tracing and tracking, the boundary between what constitutes public and private is called into question. Whilst it may bring benefits, big data also bring risks and problems as they touch almost every aspect of life (Mayer-Schönberger and Cukier, 2013; Beneito-Montagut, 2017). Using big data raises many ethical questions: privacy; traceability; ethics and accountability; surveillance; individual human agency, free will (e.g. in opting out of being traced) and responsibility; informed consent; the use and re-use of data that are stored about us; ownership of data; the threats to anonymity from re-identification of people by combining data; and the dangers of propensity analysis in judging risk and in making predictions, fair judgements and decisions about individuals (Collmann and Matei, 2016). As Mayer-Schönberger and Cukier (2013) remark, big data can 'paralyze privacy' (p. 152).

On the one hand, big data are useful (Beneito-Montagut, 2017). For example, data sets on school and student performance, attendance, grade retention and repetition, dropout, student evaluations of teaching, added value, socio-economic indicators and so on are widely used. On the other hand, this raises major questions of privacy and confidentiality, which we explore below (see also the Council for Big Data, Ethics and Society: http://bdes.datasociety.net).

It is not only in the sphere of big data that issues of ethics and privacy are raised. The Internet has spawned a raft of issues concerning privacy in research, and we introduce these here. The matter does not stop at the level of individuals. As metadata, social networking and social networking analysis are increasingly being used in research, individuals, groups, institutions and a range of other parties are caught in the debate about what constitutes legitimate and illegitimate use of electronic data. Here we focus on issues of privacy.

Solove (2004) notes that, with the rise of 'digital dossiers' and electronic data storage in many forms, the issue of privacy has come into prominence, that 'privacy is dead' (p. 73) and that people should no longer expect it in many areas that previously had been deemed private (p. 225). Indeed he quotes the CEO of Sun Microsystems as saying that there is 'zero privacy. Get over it' (p. 224) and that a new legal architecture is required to address the new, non-privacy environment. Nonetheless, privacy must still be respected, and he sets out several ways of addressing it.

Van den Hoven (1997) identifies key issues in 'privacy moral wrong-doing in an information age' (p. 33), which include: (a) the *tension between privacy, anonymity and the public good* in panoptic technologies; (b) the risk of *harm* from access to, and use of, personal information; (c) the issue of *inequality*, wherein when people use ICT they divulge not only personal information but data which are useful to organizations but to the use of which the person has not consented (van den Hoven gives an example in that each time a customer buys something they also have something to sell, namely, purchasing information (p. 35)); (d) *injustice* (e.g. discrimination and loss of agency in educational opportunity based on information from medical data stored electronically); and (e) *encroachment on moral autonomy* which occurs when privacy is compromised (or indeed shared in social networking) even with data protection laws in place.

Solove (2006) sets out a 'taxonomy of privacy' which can be applied to Internet research:

- information collection: surveillance; interrogation (probing for information);
- information processing: aggregation (combining data about a person); identification; insecurity (improper access and information leaks); secondary use (information collected for one purpose being used without consent for another purpose); exclusion (failure to inform the person that data on them is held by others and failure to involve the person in the use of such data);
- information dissemination: breach of confidentiality; disclosure (of information that affects how others judge a person's character); exposure (e.g. of bodily functions, nudity, grief); increased accessibility; blackmail (threat to disclose information); appropriation (use of a person's identity to serve the purposes or interests of another person); distortion (spreading false or misleading information about a person);
- invasion: intrusion (into a person's solitude or tranquillity); decisional interference (governmental incursion into a person's decision on private matters).

As this taxonomy indicates, the boundaries between public and private spaces are blurred in online research (e.g. Rosenberg, 2010; Brooks *et al*., 2014), and this creates challenges for informed consent. Bruckman (2004) notes that public/private spaces are not a simple binary matter – one or the other – but are a question of degree. Indeed different cultures have different conceptions of what constitutes 'public' and 'private' (Association of Internet Researchers, 2012). This links closely to the issue of informed consent. Are postings, blogs and social networking data public or private? For example, Denzin (1999) suggests that postings on bulletin boards are automatically public and so do not need informed consent for use by researchers, but is this so, and does this extend to traceability, and, if so, should not informed consent be obtained? Is the expropriation of online data for research purposes acceptable simply because it has been posted online?

How private should documents and data be, and is it ethical to use data that were not originally posted for research or public usage, for example, blogs, web pages, discussion forums, chat rooms (Denscombe, 2014, p. 321); has copyright been breached (p. 323)? Hudson and Bruckman (2005, p. 298) suggest that 'people in public, online environments often act as if these environments were private', and that they feel that their privacy has been violated if data from public chat rooms are used for research purposes, even though the data cannot not be traced back to participants.

Some data are unproblematically public, for example, national archives, publications, etc. Some may require passwords and this may require researchers to agree to 'cookies' being deposited on their computer, rendering them traceable. However, it is unclear whether other places are public or private. Is a chat room a private or a public place, or, perhaps, a semi-private space? Is an online forum a private space for members of the 'community' in question or a public space?

Rosenberg (2010) suggests that public data are those which can be accessed freely by anyone through the Internet and private if 'they are perceived as private by participants' (p. 24), in which latter case they may not be intended for public use, even if the public can access them. Like a public park, a virtual place (e.g. a cafe) can appear to be a public place, but it may be a parochial place (used by groups) or even a private place (e.g. used by clubs or parties, couples for a private conversation etc.) (cf. Rosenberg, 2010, pp. 33–4).

The researcher, then, has to decide how the participants view the space and what expectations and intentions they may have for privacy (cf. Denscombe, 2014, p. 321): public, private or somewhere in between. Simply because it is a public place does not make what happens in it completely public, in terms of both access to, and release of, information (and who is the audience of such data), and simply appropriating content because it happens to be public or accessible is ethically questionable.

Is a private communication 'fair game' for public access or researcher use, without consent? What constitutes 'private information' is blurred in Internet research. For example, it may be that in which a person can reasonably expect the context to be such that no observation, recording, monitoring or data collection is taking place or in which the individual can reasonably expect the information not to be made public (e.g. medical or financial records). However, the Internet and the tracking and searching, indeed hacking, which may accompany it pose threats to this conception of private information. Social networking sites are clear instances of this ambiguity: the data are publicly viewable, so does this mean that they are no longer private information, or only for 'friends', or for researchers, unseen or visible?

James and Busher (2007) argue that online research poses difficult issues of confirming the authenticity of respondents and responses, and of protecting the privacy of vulnerable groups, confidentiality and anonymity, particularly if emails are being used, as these are susceptible to others' viewing them either deliberately or accidentally (e.g. if mails are forwarded or shared). Further, there is a possibility that online correspondents may or may not distort their stated views, or, indeed, withhold them (p. 107), in ways that may not be so likely in face-to-face research. People may not be honest in reporting personal details; they may create avatars that have little relationship to their true selves.

Privacy and its protection include confidentiality of data and people, and this is particularly the case in sensitive Internet research or research which may bring harm or embarrassment to individuals if their identities are disclosed (and cyber-bullying and cyber-stalking are examples of this). Indeed privacy is a matter of legislation in many jurisdictions.

Privacy can be addressed by, for example, scrubbing data to remove all personal identifying material, or by providing restricted access and anonymity in the data-collection process, or by using pseudonyms, or by using encryption techniques (though some jurisdictions consider encryption to be illegal). However, as the network is not owned by, or under the control of, the researcher, scrubbing out and stripping out all potential identifiers to ensure anonymity and confidentiality may still not give an absolute guarantee that people may not be traced (Ohm, 2009). Using pseudonyms may not guarantee anonymity, not least as people use pseudonyms that describe themselves.

If researchers feel that the participants expect the data to be private, then this may raise requirements of confidentiality, anonymity, privacy and what may be made public. Just because a participant wants the data to be kept private, should those data be kept private? This raises the issue of the importance of establishing rapport and trust between the researcher and the participants. Lewis (2006), for example, took five months to establish such a relationship of trust, and he developed this relationship with his participants as a member of an online community before he approached them to participate in his research.

Whilst data protection is subject to legislation, the issue for Internet research is that, as data can be accessed from different jurisdictions, that legislation may not apply in countries outside those in which the data are generated or stored; this is a familiar issue in the protection of intellectual property.

8.5 Confidentiality and anonymity

Privacy, confidentiality and anonymity are linked in Internet research. Anonymity is where not even the researcher knows who the person is, and confidentiality is where the researcher knows but nobody else knows or is allowed to know. Researchers must consider whether, and how, to address confidentiality and anonymity in Internet research. On the one hand, not to acknowledge data sources could be deemed an infringement of copyright, even theft of intellectual property, but on the other hand such disclosure might breach participants' important right to protection from harm (Barnes, 2004).

Given that data and IP addresses are stored on networks and clouds which are not owned or controlled by the researcher but for which the researcher has a

duty of 'stewardship' (Buchanan and Zimmer, 2012), it may be impossible to guarantee anonymity and confidentiality. Indeed, combining data may relatively easily enable individuals to be 're-identified' (Ohm, 2009; Association of Internet Researchers, 2012) even in would-be anonymized data. Further, some online data-collection instruments indicate, for example, in the introductory statements to the software, that the provider owns the data, and many people do not read the small print before checking the 'I agree' or 'I accept' box. As data are held in electronic form, the software used may not permanently destroy deleted data, as 'the system' automatically keeps a digital record, which is, for example, in the permanent or semi-permanent records held by Internet companies of searches performed by individuals, or data that are entered on 'cloud' computing sites. The implications here are that it may be incorrect to promise that the data will be permanently destroyed, or, indeed, that hackers may not be able to break into the data (Marshall and Rossman, 2016, p. 182).

As it may not be possible to guarantee complete confidentiality and anonymity, and as traceability may be possible in the Internet, it is important to ensure that permissions, where required, have been given for data to be used. For example, Zimmer (2010) reports a study where, despite many precautions taken in good faith to protect ethics of consent, privacy, confidentiality, non-identifiability, personal information, non-traceability and access, including clearance by institutional review boards, nevertheless identification was uncovered easily and quickly.

In going online, there is the risk that participants may be prey to predatory Internet users. How is the protection from maleficence addressed? Researchers may need to consider how to protect participants from cyber-bullying or too-public disclosure.

A related issue in online research concerns the recruitment of participants. There is a need for researchers to establish not only their own bona fide status but that of their correspondents. This raises issues of authenticity and how to judge it (James and Busher, 2007) and the authentication of the participant's identity (a particular challenge if minors are involved, as this raises matters of legality and informed consent and the age of consent, which may differ in different jurisdictions). Further, it raises issues of anonymity, privacy and confidentiality, who is actually involved in the research (for example, is the same person completing an entire survey), or whether the research participants are operating in the environment with similar levels of control. Some software sites for research (e.g. surveys) may store cookies onto IP addresses, and this reduces the protection of privacy and confidentiality.

8.6 Ethical codes for Internet research

Many organizations have produced codes of ethics for online and Internet research. These also complement and refer to legal regulations and requirements. We give some examples below, interspersed with references to other studies on relevant ethical matters.

An early statement of research ethics, as indicated in Chapter 7, was the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979), which identified three key principles: respect for persons, beneficence and justice. Buchanan and Zimmer (2012) note that the focus of much computer security research concerns the prevention of harm to humans (p. 6). But what, exactly, is meant by a human subject (see also the Association of Internet Researchers, 2012), as the Internet, as mentioned earlier, enables people to assume different identities, to create avatars and virtual persons?

The British Psychological Society's (BPS) *Ethics Guidelines for Internet-mediated Research* (British Psychological Society, 2013) identifies four key principles (p. 5):

- respect for the autonomy and dignity of persons (covering the public/private distinction, confidentiality and anonymity, copyright, valid consent, withdrawal and debriefing);
- scientific value (including levels of control);
- social responsibility (including disrupting social structures);
- maximizing benefits and minimizing harm.

The BPS, whilst recognizing that there is a blurring of the public/private domains, and that there are differing opinions on what the boundaries of these are, those data which are readily accessible by anyone, or those data in which there is no expectation of privacy, can be considered public and no consent may be justifiable. However, where there is ambiguity, the researcher must consider the possible damaging effects on participants of undisclosed observations or those without informed consent and this may require consent, confidentiality and/or anonymity.

'Valid consent', the BPS document contends, should be obtained where it cannot be reasonably assumed that the data are public (2013, p. 8), and it recognizes that this might be problematic if the data are anonymous

(e.g. questionnaires, though completing a questionnaire might be a fair proxy for consent). Consent statements with a check box can be used, and radio buttons can be used to try to ensure that participants have read and consented (though this is no guarantee that they have read such statements). Similarly, 'exit' and 'withdraw' radio buttons can be used, and these can link to a debriefing button once the participant has completed, exited or withdrawn.

The BPS recognizes that it may be impossible to guarantee confidentiality, as researchers do not have control over the network, and email, particularly unencrypted email, is not secure. Further, as mentioned earlier, it is possible to track down an individual's IP address from forums, chat rooms, blogs, postings and verbatim quotations. Indeed researchers must ask themselves whether they need consent to use or publish such verbatim quotations. 'Consequential risk' (Williams, 2012) of harm from, say, using quotations from people should be considered, and even identifying and publishing websites might be risky to individuals or communities (p. 18).

With regard to 'scientific value', the BPS notes that researchers must address issues of control: who has access to participate; the 'environmental conditions under which the participants are responding' (2013, p. 14); participants' feelings and reactions; and variations in the research brought about by different hardware and software that the participants are using. This echoes Williams (2012), who argues that lack of controls is a serious problem for researchers, including knowing who is completing the online survey (and whether it is the same person throughout) (p. 2). Control must also be in place to prevent repeat submissions (some online survey software already builds in such checks and preventions). Similarly, with regard to maximizing benefits and minimizing harm, the researcher must take steps to ensure the protection of minors (e.g. in informed consent) and verifying identity. Williams (2012) notes that often it is teenagers who not only use social networking most but are most at risk from being harmed and traced by it, not least because they may not realize that they are being monitored unobtrusively; they may also be subject to cyber-bullying (pp. 3–4). It may be necessary to avoid situations where researcher controls are so few that there is a real risk of harm to participants.

'Social responsibility' concerns beneficence and the betterment of society. This extends to covert research, and Orton-Johnson (2010) notes that it is relatively easy to conduct covert research on the Internet. The researcher has to decide whether to use covert research (non-disclosure that he/she is a researcher, see Chapter 7). For example, some researchers may join online communities without disclosing the fact that, actually, they are doing so in order to obtain research data (Reynolds and de Zwart, 2010). Is this ethical? Should they disclose their intentions and seek permission to participate and make data public?

'Maximizing benefits and minimizing harm' requires a risk assessment and an identification of the nature, duration, degree, severity, intensity, discomfort of risk and harm (physical, emotional, psychological, social etc.), how to address it, and how to balance it with possible benefits. We refer readers to the considerable discussion of this in Chapter 7. As mentioned in Chapter 7, the research must not leave participants worse off at the end of the research than they were at the beginning (non-maleficence); indeed maybe their own and others' lives should have been improved by participation (beneficence).

Ess and the Association of Internet Researchers (2002) set out ethical guidelines for researchers using the Internet for data collection and research, including:

- Do not assume that emails are secure.
- Ensure that nobody is harmed by the research.
- Enable participants to correspond in private if they wish.
- Indicate the steps taken to ensure privacy.
- Check where the communication comes from.
- Determine the most suitable online method of requesting and receiving informed consent.
- The greater the acknowledged publicity of the venue, the less obligation there may be to protect individual privacy, confidentiality and rights to informed consent.
- The greater is the vulnerability of the researcher to the participant, the greater is the obligation of the researcher to protect the participant.
- Indicate clearly how material will be used and whether or how it will be attributed, and whether data will be used verbatim, aggregated or summarized.
- Work within the framework of legal obligations of protection (e.g. data protection, privacy, copyright and libel laws).
- Indicate who has access to the communication, and whether it is private.
- Consider the possible outcomes to individuals if private data are made public.

Similarly, Gwartney (2007) argues for professional ethics to be respected, and she indicates websites that can provide guidance to researchers on this (p. 53), including codes of conduct, informed consent,

confidentiality, privacy, avoidance of harassment, email solicitation, active agent technology (e.g. behind-the-scenes data mining), installing software and setting cookies or hard-to-uninstall software, codes and standards for minimal disclosure, unsolicited telephone calls and setting up 'Do Not Call' lists, professional responsibilities in working with people. Additionally, Gwartney (2007) reproduces some of these ethical guidelines (pp. 57–69).

More recently, the Association of Internet Researchers (2012) provides a comprehensive set of guiding principles for ethical Internet research. These recognize the situated, contextual nature of ethical decision making, such that there may be no single set of judgements (no 'one-size-fits-all'; p. 4) which is universally applicable. Rather, researchers have to take ethical decisions on a case-by-case, casuistic basis (p. 7) (see the discussion of this in Chapter 7). The Association's 'key guiding principles' include:

- The greater the vulnerability of the community/author/participant, the greater the obligation of the researcher to protect the community/author/participant.
- Because all digital information at some point involves individual persons, consideration of principles related to research on human subjects may be necessary even if it is not immediately apparent how and where persons are involved in the research data.
- When making ethical decisions, researchers must balance the rights of subjects (as authors, as research participants, as people) with the social benefits of research and researchers' rights to conduct research. In different contexts the rights of subjects may outweigh the benefits of research.
- Ethical issues may arise and need to be addressed during all steps of the research process, from planning, research conduct, publication, and dissemination.
- Ethical decision-making is a deliberative process, and researchers should consult as many people and resources as possible in this process, including fellow researchers, people participating in or familiar with contexts/sites being studied, research review boards, ethics guidelines, published scholarship (within one's discipline but also in other disciplines), and, where applicable, legal precedent.

(Association of Internet Researchers, 2012, pp. 4–5)

The Association also recognizes that key considerations of potential harm, vulnerability, beneficence and respect for people apply throughout the research process (2012, p. 5). In keeping with its advocacy of a case-by-case approach to ethics, the Association raises some eighty questions that researchers can address in considering the ethics of their Internet research, including:

- How is the context defined?
- How is the content (venue/participants/data) being accessed?
- Who is involved in the study?
- What is the primary object of study?
- How are data managed, stored, and represented?
- How are texts/persons/data being studied?
- How are findings presented?
- What are the potential harms or risks associated with this study?
- What are the potential benefits associated with this study?
- How are we recognizing the autonomy of others and acknowledging that they are of equal worth to ourselves and should be treated so?
- What particular issues might arise around the issue of minors or vulnerable persons?

(Association of Internet Researchers, 2012, pp. 8–11)

The Association's document also provides a useful chart of types of data, venues and contexts, and commonly asked questions concerning ethical practice.

The UK's Economic and Social Research Council (2015) argues that risk assessment must include research involving 'social media and participants recruited or identified through the Internet, in particular when the understanding of privacy in these settings is contentious where sensitive issues are discussed' (p. 10). Further, whilst it defines differences between public and private domains, the former being those 'forums or spaces on the Internet that are intentionally public' (p. 12), it also argues that

> the public nature of any communication or information on the Internet or through social media should always be critically examined, and the identity of individuals protected, wherever possible, unless it is critical to the research, such as statements by public officials.
>
> (p. 12)

It also notes that social media users must abide by any regulations set out by those social media and data providers, and it offers a cautionary note that children and others 'may not understand the implications of what they are doing, and those harvesting data may also uncover illegal images or activities' (p. 12). Researchers, they comment, must consider issues of anonymity

in social media and place themselves in the shoes of the participants in considering whether the data from social media are in the public or private domains (p. 26).

8.7 Conclusion

An overriding principle is the double issue of non-maleficence and beneficence. It is easy to use the phrase 'do no harm'. However, as seen in this chapter and in Chapter 7, it is neither easy to define nor easy in practice, particularly where individual privacy may conflict with the public good. We have also noted the importance of addressing legal requirements and constraints. Further, this chapter has suggested that it is important for Internet researchers to take defensible decisions on many issues (e.g. Watson *et al*., 2007; Association of Internet Researchers, 2012; British Psychological Society, 2013), for example:

- decide whether the participants themselves consider the virtual community to be a public or private space and online data to be public or private, and to what degree. This might be informed by consideration of membership and membership access (e.g. whether it is open or restricted, a private, intimate group, stable membership). Give serious consideration to participants' expectations and perceptions;
- decide how to respect the autonomy and dignity of individuals;
- decide how to ensure the scientific value and control of the online research;
- decide whether or how much the research is overt, covert, obtrusive, unobtrusive, intrusive or non-intrusive, socially disruptive or non-disruptive, and justify the decision;
- decide the ethics of access to people and data (e.g. covert, overt, deception, intrusion, non-intrusion, intrusiveness, unobtrusiveness);
- decide whether and how to verify authenticity and identity;
- decide how to address privacy, confidentiality, anonymity and non-traceability;
- decide on removal of identifying data;
- decide the vulnerability of the group and the potential risk of harm to the participants (including minors and vulnerable people);
- decide how to address non-maleficence, beneficence and the minimization of harm;
- decide whether informed consent is required. If so, from whom, when, for what (from access to publication), for how long (including archived data), what constitutes 'valid consent' and 'informed', and how the consent will be obtained. If informed consent is not required, then such a decision must be defensible;
- decide how to address data removal if participants withdraw;
- decide how to address debriefing;
- decide how to establish a relation of trust with online contacts where appropriate (e.g. in ethnographic research);
- decide who owns the data, and for how long, and what are the intellectual property rights and responsibilities;
- decide how data will be stored and archived securely, and with what protections;
- decide how to report, disclose and disseminate the research ethically, with appropriate protections.

As can be seen, many of the issues listed above rehearse ethical challenges in everyday research, i.e. they are not exclusive to Internet and online research. However, careful attention needs to be given to these and how they are applied and interpreted in online research.

Companion Website

The companion website to the book provides PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Choosing a research project

CHAPTER 9

This chapter provides key decision points of reference on which researchers can reflect and plan, including:

- how to choose a research project
- the importance of the research
- the purposes of the research
- ensuring that the research can be conducted
- research questions
- the scope of the literature review
- a summary of key issues in choosing a research topic or project

This chapter concerns the selection of the research and initial, practical matters that researchers can address when choosing and deciding the project on which to work. It is the first of six consecutive chapters that concern the planning of research. This chapter concerns the selection of the research and the initial matters to address, whilst the subsequent chapters unpack several of these in greater detail. We draw not only from relevant literature but from our own experiences of supervising several hundred research students. Research is a practical activity, and the advice that we give here is practical. This is not a simplistic recipe or low-level 'tips for researchers'; rather it is the distillation of key features of practicable research and issues on which to deliberate, and to help to ensure that the research provides relevant and useful findings.

9.1 Introduction

Choosing a research project is normally the decisive feature of successful research. Many novice students and researchers start with an over-ambitious project. The task of a mentor or supervisor is to help the novice researcher to narrow and hone down the research field in order to render the research practicable, useful and workable. Indeed part of the discipline of choosing and conducting a piece of research is fining it down to manageable/researchable proportions, to enable rigour (e.g. fitness for purposes and methodological soundness) to be inserted into the research. Rigour in planning and doing research lies in choosing a project that is sufficiently tightly framed. A research topic is only one small aspect of the field of the subject, and careful boundaries must be drawn around the topic: what it will and will not do.

For novice researchers, a piece of educational research often starts by wanting to be their life story or the opportunity to give their personal opinions some grounding in literature and empirical study that support their opinions or prejudices. This is not the task of research. The task of research is to find out, to investigate, to develop, to test out (e.g. a theory), to address questions such as: 'what if', 'how', 'why', 'how well', 'what' and 'where'.

9.2 What gives rise to the research project?

Several points can give rise to a research topic. For example, for many teachers it may be a problem that they encounter in their day-to-day work: they may want to find out the causes of the problem and how to solve it; they may want to plan an intervention to see how well it addresses or solves the problem. Examples of these might be: 'How can teachers improve students' learning of algebra in lower secondary schools?'; 'How to maximize the learning of students with Asperger's syndrome in mainstream schooling'; 'How to conduct a music lesson with many musical instruments, without the lesson descending into chaos and noise'; 'How to teach speaking a foreign language in large, mixed-ability classes'.

Some research projects may begin with an area of interest or personal experience that researchers may have been wanting to investigate, for example: 'What is the long-term effect on employment of early school dropout?'; 'How effective is early identification of behaviour disorders on educational provision for such students?'; 'How can teachers improve students' motivation to learn a second language?'; 'Why do young teachers leave teaching and older teachers stay?'

Some research topics may begin with a recognized area of importance or topical concern in the field, for example: 'How to maximize primary students' learning

using ICT'; 'What is the effect of frequent testing on students' stress?'; 'How can developments in brain research and cognitive neuroscience impact on pedagogy?'; 'What is the predictive validity of personality tests or learning styles inventories on the success of first-time employees' applications for employment?'; 'Do interactive teaching methods produce higher test scores in university students than lecture-based teaching?' Such importance may arise from coverage of the topic in the press, articles, conference papers and journals.

Some research is conducted as part of a sponsored research project, in which the field and purposes of the research must be spelled out very clearly in order for the sponsorship to be obtained. For example in the UK the Economic and Social Research Council (www.esrc.ac.uk/research/research-topics), the Leverhulme Trust (www.leverhulme.ac.uk), Nuffield Foundation (www.nuffieldfoundation.org) and the Joseph Rowntree Foundation (www.jrf.org.uk) require detailed applications to be completed, and in the United States the Social Science Research Council (www.ssrc.org) requires similarly high levels of detail. Such funding might also need to fit the categories of research set out by the funding agencies.

A decision on what to research can arise from several wellsprings of the researcher's own motivation:

- a problem encountered in the researcher's everyday work or outside her/his everyday work (e.g. conceptual, theoretical, substantive, practical, methodological);
- an issue that the researcher has read about in a journal, book or other media;
- a problem that has arisen in the locality, perhaps in response to government policy or practices or to local developments;
- an area of the researcher's own interest;
- an area of the researcher's own experience;
- a perceived area of importance;
- an interesting question;
- a testable guess or hunch;
- a topical matter;
- disquiet with a particular research finding that one has met in the literature or a piece of policy (e.g. from the school, from a government), and a wish to explore it further;
- an awareness that a particular issue or area has been covered only partially or selectively in the literature, and a wish to plug the gap;
- a wish to apply a piece of conceptual research to actual practice, or to test a theory in practice;
- a wish to rework the conceptual or theoretical frameworks that are often used in a specific area;
- a wish to revise or replace the methodologies that are often used in researching a specific area;
- a desire to improve practice in a particular area;
- a desire to involve participants in research and development;
- a desire to test out a particular methodology in research;
- an interest in seeing if reported practice (e.g. in the literature) holds true for the researcher's own context (e.g. a comparative study);
- an interest in investigating the causes of a phenomenon or the effects of a particular intervention in the area of the phenomenon;
- a wish to address an issue or topic that has been under-researched in the literature;
- a priority identified by funding agencies;
- an issue identified by the researcher's supervisor or a project team of which the researcher is a member;
- a wish to explore further or to apply an issue or topic that one has encountered, for example, in the literature.

The long list above concerns the motivation that leads a researcher to consider doing a particular piece of research. Add to this a salutary point for researchers, which is that the study on which they might embark will probably take weeks, months and maybe years. Sustaining interest and momentum in the researcher(s) are important considerations. Researchers should ask themselves whether they really have the interest in studying the issue in question or in conducting the research for a long period of time. If the answer is 'no' then, if they have the luxury of not having to do this particular piece of research, they may wish to consider an alternative area that will enable them to sustain interest in, and motivation for, the research. A piece of research that is conducted by an unwilling or bored researcher could easily become unimpressive.

Beyond the *motivation* for the research are the *sources* of the research in question: where does research come from? For example, the research may derive from:

- a practical concern (e.g. 'why do females have higher scores than males in international tests of reading at age 14?') or a practical need (Leong *et al*., 2012);
- a literature review (though Andrews (2003) observes that if the research question derives from the literature review then there is a risk that there is no research question to initially drive the literature review (p. 18), i.e. the literature review could lack direction, purpose and boundaries). A literature

search (including specialist literature in the field, primary and secondary sources) helps the researcher to understand the existing field and the real-world implications of the research (Alvesson and Sandberg, 2013);
- the identification of a gap in the literature or field of study (gap filling) (Alvesson and Sandberg, 2011, 2013);
- the identification of where the research can build on existing literature (Alvesson and Sandberg, 2011);
- a theoretical concern, enabling theories to be generated and tested (e.g. 'how significant is performance-related pay in motivating senior managers of schools?', in which the 'theory' to be tested is that performance-related pay is a necessary but not sufficient motivator of senior staff (Pink, 2011));
- policy concerns (e.g. 'how effective is such-and-such in attracting females to take STEM subjects?');
- concerns in the media and blogs (including the Internet);
- society, empirical data (Alvesson and Sandberg, 2013, p. 16);
- personal experience, interest or observation (Leong *et al*., 2012);
- colleagues and contacts (ibid.);
- experts and practitioners in the field (ibid.);
- conferences and conventions (ibid.);
- faculty seminars, research groups, discussion groups and workshops (ibid.);
- students (ibid.);
- societies, associations, research bodies and special interest groups;
- spotting where areas are neglected, for example, overlooked/under-researched;
- existing studies and influential theories (Alvesson and Sandberg, 2013, p. 17);
- challenge to, or problematization of, an assumption, agenda or existing theory (Alvesson and Sandberg, 2013);
- a novel idea which challenges existing ideas or practices;
- funding bodies and/or project directors;
- spotting where applications may lie;
- spotting where confusions need to be clarified;
- spotting where new methodologies and research methods might be applied;
- other starting points – the list is endless.

It is essential that the research and the questions it asks should address something that is worth asking: asking the right question (Leong *et al*., 2012, p. 121). In turn this means that the research itself must be worth doing – it must make a significant contribution to the field. Behind the many features of effective research questions lies the need to ensure that the research itself, i.e. in principle, is interesting. In this respect there is an overlap in the literature between research areas and research questions, i.e. what some authors would place under the category of 'research questions' could just as easily be placed in the category of 'research areas' or 'fields of research', or 'research topics'. This harks back to the seminal work of Davis (1971) (see also Chapter 4), who provides a formidable list of twelve factors that make social science, and hence research and research questions, 'interesting'.

More recently, Alvesson and Sandberg (2011, 2013) argue that much research is 'gap filling', and that, whilst worthy, this risks being over-confined to the status quo, conservative, under-problematizing or over-problematizing matters, derivative and non-interesting because, since it builds on or around existing literature, it does not challenge assumptions in the literature, does not sufficiently problematize assumptions and agendas, and does not generate really new ideas or innovatory, creative thinking. It reinforces rather than challenges consensus (Alvesson and Sandberg, 2011, p. 250). Gap spotting, they observe, might be easy, uncontroversial and resonant with the idea of cumulative research, but it does not question received wisdoms and research perspectives.

Rather, Alvesson and Sandberg (2011, 2013) argue for the problematization of issues and the development of new ideas – *challenging* assumptions, agendas and theories – in order to create 'interesting' and 'influential' research and research questions (2013, p. 45). Problematization and questioning assumptions, they suggest, is a powerful methodology for generating interesting research questions and questioning of received truths, i.e. disruptive of existing theory, practices, paradigms and ideologies, and it is faithful to the uncertain nature of scientific 'truths' (p. 50). The aim of problematization, they argue, is to 'disrupt rather than build upon and extend an established body of literature' (2011, p. 248).

Of course, gap filling, building on existing research and problematization for the creation of new ideas are not mutually exclusive. All can generate 'interesting' research; as the authors remark (Alvesson and Sandberg, 2011, p. 266), there are good reasons for gap spotting as this can enable research to supplement and enrich existing studies, and clarify issues, for example, where there are disagreements among researchers. Innovative, high-impact research questions, they suggest (Alvesson and Sandberg, 2011, 2013), stem from the questioning of assumptions that underlie existing theories in significant ways. They set out a methodology

for problematization to produce 'interesting' research and research questions which constitutes one of Davis's (1971) features of 'interesting' research: what appear to be matters or phenomena that can coexist actually cannot, and vice versa (p. 4). Alvesson's and Sandberg's (2011, p. 256) methodology for generating 'interesting' research through 'dialectical interrogation' of assumptions requires researchers to:

Step 1: Identify a domain of literature;
Step 2: Identify and articulate the assumptions that underlie that domain;
Step 3: Evaluate the assumptions that underlie that domain;
Step 4: Develop an alternative assumption ground;
Step 5: Consider this alternative assumption ground in relation to its audience;
Step 6: Evaluate the alternative assumption ground.

Essentially the task is to expose and evaluate existing 'in-house' assumptions (e.g. in the literature, in 'theories'), i.e. those assumptions which are regarded as unproblematic and which are accepted by their advocates (p. 254), thence to challenge those assumptions (e.g. problems with them, their shortcomings and oversights) (p. 267), and develop and evaluate an 'alternative assumption ground' that will generate 'interesting' theory, taking the latter into account in relation to the audience, i.e. the wider intellectual, social and political situation of the research community and their possible reactions to the challenges posed (p. 258), and check to see if the alternative assumption ground is obvious, interesting or, indeed, absurd (p. 259).

Alvesson and Sandberg argue, for example, that rather than trying to develop research and research questions solely from a literature review, it might be more 'interesting' (and they use Davis's (1971) word here) to ask how a particular field becomes the target of investigation, to evaluate and challenge the *assumptions* (unchallenged, accepted and shared schools of thought), *ideologies* (e.g. values, politics, interests, identifications, moral and ethical views), *paradigms* (ontological, epistemological and methodological assumptions, world views), *root metaphors* (images of a particular area) and *field assumptions* (broader sets of assumptions about specific subject matter which are shared by schools of thought within, across a paradigm or discipline) (2011, p. 255) that underlie a theory. From there, the researcher seeks to develop and evaluate the 'alternative assumption ground' which, thereby, is 'more disruptive' and 'less reproductive' (Alvesson and Sandberg, 2013, p. 122). Challenging in-house assumptions is regarded as a minor level of problematization (Alvesson and Sandberg, 2011, p. 255); questioning root metaphors constitutes a middle-ground challenge; and challenging ideology, paradigms and field assumptions constitutes a more fundamental form of problematization (p. 255).

Leong *et al*. (2012, pp. 128–9) suggest that research and its research questions can be framed which: (i) discover a new effect; (ii) extend an established effect (e.g. to new domains); (iii) demonstrate mediation of factors (interaction), i.e. the mechanisms that lead to an effect; and (iv) moderation of an established effect (modelling for which groups of people/situations the effects hold true or not true). Whilst discovering a new effect may be for seasoned researchers, they note that extending an established effect may be suitable for novice researchers. They comment that moving beyond 'gap filling' to novel research is uncomfortable because it takes us out of our familiar, sedimented, deeply ingrained ways of thinking. They suggest that making the opposite assumptions, exposing hidden assumptions, casting doubt on existing assumptions and scrutinizing meanings of key concepts is unsettling (pp. 126–7).

Alvesson and Sandberg (2011, 2013) are arguing that effective, high-impact research and research questions derive from high-impact research proposals that move beyond 'gap filling' to disrupting conventions, modes of thinking and examining a phenomenon. This echoes Leong *et al*. (2012) who argue that creative, innovative, worthwhile research may be unclear at the outset and that if it is too clear too early on then it may not be focusing on anything new or important (p. 122); as the authors say, if it is too predictable, why do it? Indeed they write that an innovative research question is one that generates ambiguity rather than certainty, and they suggest that effective research questions are those which: are unclear on their outcomes; can generate answers; and discriminate between theories, each of which leads to different predictions (p. 122).

9.3 The importance of the research

Whatever research area or topic is identified, it is important for it to be original, significant, non-trivial, relevant, topical, interesting to a wider audience and to advance the field. For example, I may want to investigate the use of such-and-such a textbook in Business Studies with sixteen-year-olds in Madagascar, but, really, is this actually a useful research topic or one that will actually help or benefit other teachers or educationists, even though it yields original data?

Or I might conduct research that finds that older primary children in a deprived area of Aberdeen,

Scotland prefer to have their lunch between 12 noon and 1.00 p.m. rather than between 1.00 p.m. and 2.00 p.m., but, really, does anybody actually care? The topic is original and, indeed, the data are original, but both are insignificant and maybe not worth knowing.

In both of these examples, the research brings about original data, but that is all. Research needs to go beyond this, to choose a significant topic that will actually make an important contribution to our understanding and to practice. Originality alone is not enough. Rather, the research should move the field forward, perhaps in only a small-scale, piecemeal, incremental way, but nevertheless to advance it such that, without the research, the field would be poorer. Hence it is important to consider how the research takes the field forwards not only in terms of data, but also conceptually, theoretically, substantively and/or methodologically. At issue here is not only the contribution to knowledge that the research makes, but the *impact* of that knowledge; indeed funding agencies typically require an indication of the impact that the research will make on the research community and more widely, and how that impact will be assessed and known. What will be the impact, uptake and effects of the research, and on whom?

It is also useful for the researcher to identify what benefit the research will bring, and to whom, as this helps to focus the research and its audience. Fundamental questions are 'what is the use of this research?' 'What is the point of doing this research?' 'Who benefits?' 'Is this research worth doing?' If the answer to the last question is 'no', then the researcher should abandon it, otherwise it ceases to be useful research and becomes an indulgence of the dilettante.

Many novice researchers may not know whether the research is original, significant, important, complex, difficult, topical and so on. Here it is important for such a novice to read around the topic, to conduct a literature search, to conduct an online search, to attend conferences on the topic, to read newspaper reports on the topic; in short, to review the state of the field before coming to a firm decision on whether to pursue research in that field. In this respect, if the researcher is a student, it is vital to discuss the proposed topic with a possible supervisor, to receive expert feedback on the possible topic.

Before a researcher takes a final decision on whether to pursue a particular piece of research, it is useful to consider selecting a topic that interests the researcher, reading through background materials and information and compiling a list of keywords, clarifying the main concepts and writing the topic as a statement (or a hypothesis). Whilst incomplete, nevertheless this provides a useful starting point for novice researchers contemplating what to research.

9.4 The purposes of the research

Implicit in the previous section is the question 'why do the research?' This is ambiguous, as 'why' can refer to reasons/causes and purposes, though the two may overlap. Whereas the previous section concerned reasons, this section concerns purposes: what we want the research to achieve. It is vital that the researcher knows what she or he wants the research to 'deliver', i.e. to answer the question 'what are the "deliverables" in the research?' In other words, what do we want to know as a result of the research that we did not know before the research commenced? What do we want the research to do? What do we want the research to find out (which is not the same as what we want the results to be: we cannot predict the outcome, as this would be to 'fix' the research; rather, the kind of information or answers we want the research to provide)?

In this respect it is important for the researcher to be very clear on the purposes of the research, for example:

- to demonstrate that such-and-such works under a specified set of conditions or in a particular context (experiment; action research);
- to increase understanding and knowledge of learning theories (literature-based research);
- to identify common features of successful schools (research synthesis; descriptive research);
- to examine the effects of early musical tuition on general intelligence (meta-analysis; multilevel research);
- to develop and evaluate community education in rural and dispersed communities (participatory research; evaluative research; action research);
- to collect opinions on a particular educational proposal (survey);
- to examine teacher–student interactions in a language programme (ethnography; observational research);
- to investigate the organizational culture of the science faculty in a university (ethnography; survey);
- to identify the relative strengths of a range of specified factors on secondary school student motivations for learning (survey; observational study; multiple regression analysis; structural equation modelling);
- to see which of two approaches to teaching music results in the most effective learning (comparative study; experiment; causal research);
- to see what happens if a particular intervention in setting homework is introduced (experiment; action research; causal research);

- to investigate trends in social networking in foreign language teacher communities (network analysis);
- to identify key ways in which teachers in a large secondary school view the leadership of the senior staff of the school (personal constructs; accounts; survey);
- to interrogate government policy on promotion criteria in schools (ideology critique; feminist critique);
- to see the effects of assigning each student to a mentor in a university (survey; case study; causal research);
- to examine the long-term effects of early student dropout from school (survey; causal or correlational research);
- to see if repeating a year at school improves student performance (survey; generalization; causal or correlational research);
- to chart the effects of counselling disruptive students in a secondary class (case study; causal or correlational research);
- to see which catches richer survey data on student drug usage: questionnaires or face-to-face interviews (testing instrumentation; methodology-related research);
- to examine the cues that teachers give to students in question-and-answer classroom episodes (discourse analysis);
- to investigate vandalism in schools (covert research; informer-based research);
- to investigate whether case studies or surveys are more effective in investigating truancy in primary school (comparative methodology);
- to run a role-play exercise on communication between a school principal and senior teachers (role-play);
- to examine the effects of resource allocations to under-performing schools (ideology critique; case study; survey; causal research);
- to understand the dynamics of power in primary classrooms (ethnography; interpretive research);
- to investigate the demise of the private school system in such-and-such a town at the end of the nineteenth century (historical research);
- to understand the nature of trauma and its treatment on primary-aged children living in violent households (case study; action research; grounded theory; ex post facto research);
- to generate a theory of effective use of textbooks in secondary school physics teaching (grounded theory);
- to clarify the concept of 'the stereotype activation effect' for investigating the effect of sex stereotyping on reading in young teenagers (survey; case study; experiment; causal research);
- to test the hypothesis/theory that increasing rewards loses effect on students over time (experiment; survey; longitudinal research; causal or correlational research).

As can be seen in these examples, different purposes suggest different approaches, so 'fitness for purpose' takes on importance in planning research (see Chapter 10). One can also see that there is a range of purposes and types of research in education. The researcher cannot simply say that he or she likes questionnaires, or is afraid of numbers, or prefers to conduct interviews, or feels that it is wrong to undertake covert research so no covert research will be done. That is to have the tail wagging the dog. Rather, the research purposes determine what follow in respect of the kind of research, the research questions, the research design, the instruments for data collection, the sampling, whether the research is overt or covert (the ethics of research), the scope of the research, and so on.

9.5 Ensuring that the research can be conducted

Many novice researchers, with the innocence and optimism of ignorance, may believe that whatever they want to do can actually be done. This is very far from the case. There is often a significant gulf between what researchers want to do and what actually turns out to be what they can do.

A formidable issue to be faced here is one of *access*. Many new researchers fondly imagine that they will be granted access to schools, teachers, students, parents, difficult children, students receiving therapy, truants, dropouts, high performers, star teachers and so on. This is usually NOT the case: gaining access to people and institutions is one of the most difficult tasks for any researcher, particularly if the research is in any way sensitive (see Chapter 13). Access problems can kill the research, or can distort or change the original plans for the research.

It is difficult to overstate the importance of researchers doing their homework before planning the research in any detail, to see if it is actually feasible to gain access to the research sites or people they seek. If the answer is 'no' then the research plan either stops or has to be modified. It is not uncommon for the researcher to approach organizations (schools, colleges, universities, government departments) with some initial, outline plans of the research, to see if there is a possibility, likelihood or little or no chance of doing the research.

Nor is it enough to be clear on access; supplementary to this is 'access to what?'. It is of little use to be given access to a school by the school principal if the teachers have not been consulted about this, or if they are entirely uncooperative (see the discussion of informed consent in Chapter 7). One of the authors recalls an example of a Master's student who wanted to study truancy; the student had the permission of the school principal and turned up on the day to commence the research with the school truants, only to find that they had truanted, and were not present! The same is true for sensitive research. For example, let us suppose that one wished to research child abuse in primary school students. The last people to consent, or even to be identified and found, might be the child abusers or the abused children; even if they *were* identified and found, why should they agree to being interviewed by a stranger who is conducting research? Or, let us suppose that one wished to investigate the effects on teachers of working with HIV-positive children in hospital; those teachers might be so traumatized or emotionally exhausted at the end of a day's work that the last thing they want to do is to talk about it further with an outside researcher whom they have never met before; they simply want to go home and 'switch off'. These are real issues. The researcher has to check out the situation before embarking on a fully worked-out plan, because the plan might come to nothing if access is not possible.

It is not only the people with whom the researcher is working who have to be considered; it is the researcher herself/himself. For example, does the researcher have the right personality, dispositions, sympathies, interpersonal skills, empathy, emotional intelligence, perseverance and so on to conduct the research? For instance, it would likely be a disaster if a researcher were conducting a piece of research on student depression and tacitly believed that students were just lazy or work-shy and that they used 'feeling down' (as the researcher might put it) as an excuse, i.e. the researcher refused to recognize the seriousness of depression as a clinical condition or as a pathological disorder. Equally, it would be an unwise researcher who would choose to conduct a longitudinal study if she had limited perseverance or if she knew that she was going to move overseas in the near future.

Researchers themselves will also need to decide whether they have sufficient expertise in the field in which they want to do the research. It could be dangerous to the researcher and to the participants if the researcher were comparatively ignorant of the field of the proposed research, as this could mean that direction, relevance, prioritization or even safety might be jeopardized. This is a prime reason for the need for researchers to conduct a literature review, to demonstrate that they are sufficiently well-versed in the field to know what to do, what to look for, and where, when and how to proceed.

Researchers will also have a personal commitment to the research; it may help to further their specialist interest or expertise; it may help to establish their reputation; it may make for career advancement or professional development. These considerations, though secondary, perhaps in choosing a piece of research, nevertheless are important features, given the commitment of time and effort that the research will require.

In addition to access, there are issues of time to be considered. Part of the initial discipline of doing research is to choose a project that is manageable – can actually be done – within the time frames that the researcher has at her/his disposal. It would be ridiculous for a researcher to propose a longitudinal study if that researcher only has maybe six or nine months to plan, conduct and report the entire research project. The time frames may prevent certain types of research from being conducted.

Similarly, the time availability of the researcher has to be considered: many researchers are part-time students who may not have much time to conduct research, and often their research is a lonely, one-person affair rather than a group affair with a team of full-time researchers. This places a practical boundary around what can and cannot be done in the research. Again, these are real issues. The availability of the researcher features in ensuring that the research can be conducted, and this applies equally to the participants: are they willing and able to give up their time in participating in the research, for example, in being interviewed, in keeping diaries, attending follow-up debriefings, participating in focus groups and writing reports of their activities?

Whilst access and time are important factors, so are resources (e.g. human, material). For example, if one is conducting a postal survey there are costs for printing, distribution, mail-back returns and follow-up reminders. If one is conducting a questionnaire survey on a large, dispersed university campus then one will need the cooperation of academic and administrative staff to arrange for the distribution, collection and return of the questionnaires. If one is conducting an online survey of teachers' views of, for example, government assessment policy, can it be assured that all teachers will have access to the online facilities at times that are convenient for them, and that poor connectivity, slow speed and instability of the system will not end in them abandoning the survey before it is completed?

If one is conducting an analysis of trends in public education in early-twentieth-century Scotland, then one needs to have time to search and retrieve public records (and this may involve payment), maybe visit geographically dispersed archives, and sit in front of microfiche readers or computers in public record offices and libraries.

A further consideration in weighing up the practicalities of the research is whether, in fact, the research will make any difference. This is particularly true in participatory research. Researchers may wish to think twice before tackling issues about which they can do nothing or over which they may exert little or no influence, such as changing an education or schooling system, changing the timetabling or the catchment of a school, changing the uses made of textbooks by senior staff, changing a national or school-level assessment system. This is not to say that such research cannot or should not be done; rather it is to ask whether the researcher's own investigation can do this, and, if not, then what the purposes of the research really are or can be.

Many researchers who are contemplating empirical enquiries will be studying for a degree. It is important that they will be able to receive expert, informed supervision for their research topic. Indeed, in many universities a research proposal will be turned down if the university feels that it is unable to supervise the research sufficiently. This will require the student researcher to check out whether his/her topic can be supervised properly by a member of the staff with suitable expertise, and, indeed, many students find this out before even registering with a particular university. It is a sound principle.

A final feature of practicality is the scope of the research. This returns to the opening remarks of this chapter, concerning the need to narrow down the field of the study. We advise that a single piece of research be narrow and limited in scope in order to achieve manageability as well as rigour. As the saying goes, 'the best way to eat an elephant is one bite at a time'! Researchers must put clear, perceptible, realistic, fair and manageable boundaries round their research. If this cannot be done straightforwardly then maybe the researcher should reconsider whether to proceed with the planned enterprise, as uncontrolled research may wander everywhere and actually arrive nowhere. Part of the discipline of research is to set its boundaries clearly and unequivocally. In choosing a piece of research, the manageability of setting boundaries is important; if these cannot be set, then the question is raised of the utility of the proposed endeavour.

For example, if one were to investigate students' motivations for learning, say, biology, this would involve not only identifying a vast range of independent variables, but also handling likely data overload, and ensuring that all the theories of motivation were included in the research. This quickly goes out of control and becomes an impossible task. Rather, one or two theories of motivation might be addressed, within a restricted, given range of specified independent variables (unless, of course, the research was genuinely exploratory), and with students of a particular age range or kind of experience of biology.

Small samples, narrowly focused research, can yield remarkable results. For example, Axline's *Dibs in Search of Self* (1964) study of the restorative and therapeutic effects of play therapy focused on one child, and Piaget's (1932) seminal theory of moral development, in *The Moral Judgement of the Child*, focused on a handful of children. In both cases, the detailed carefully bounded research yielded great benefits for educationists.

Practical issues, such as those mentioned here, often attenuate what can be done in research. They are real issues. The researcher is advised to consider carefully the practicability of the research before embarking on a lost cause in trying to conduct a study that is doomed from the very start because insufficient attention has been paid to practical constraints and issues.

9.6 Considering research questions

The move from the aims and purposes of a piece of educational research to the framing of research questions – the process of operationalization of the research – is typically not straightforward, but an iterative process. The construction of careful research questions is crucial and we devote an entire chapter to this (Chapter 10). We refer the reader to that chapter and indicate in it that research questions typically drive and steer much research.

It is the answers to the research questions that can provide some of the 'deliverables' referred to earlier in the present chapter. A useful way of deciding whether to pursue a particular study is the clarity and ease in which research questions can be conceived and answered. As mentioned in more detail in Chapter 10, research questions turn a general purpose or aim into specific questions to which specific, data-driven, concrete answers can be given. Questions such as 'what is happening?', 'what has happened?', 'what might/will/should happen?' open up the field of research questions. Chapter 6 also mentioned causal questions; here 'what are the effects of such-and-such a cause?' and 'what are the causes of such-and-such an effect?' are two such questions, to which can be added the frequently

used questions 'how?' and 'why?'. These questions ask for explanations as well as reasons.

As we mention in Chapter 10, the research may have one research question or several. Andrews (2003, p. 26) suggests that the research should have only one main research question and several supporting questions: 'subsidiary' questions which derive from and are necessary, contributory questions to the main research question (see Chapter 10 of the present volume). He notes that it is essential for the researcher to identify what is the main question and how the subsidiary questions relate to it. For example, he suggests that a straightforward method is to put each research question onto a separate strip of paper and then move the strips around until the researcher is happy with the relationship between them as indicated in the sequence of the strips (p. 39). This implies that the criteria for identifying the relationship have to be clear in the researcher's mind (e.g. logical/chronological/psychological, general to specific, which questions are subsumed by or subsidiary/subordinate/superordinate to others, which questions are definitional, descriptive, explanatory, causal, methodological etc., which question emerges as the main question). This process, he notes (p. 41), also enables the researcher to identify irrelevant questions and to refine down, to delimit the research; many novice researchers may have many research questions, each of which merits its own substantial research in itself, i.e. the research questions are unrealistically ambitious.

Chapters 1 and 2 drew attention to numerical, non-numerical and mixed methods research questions. Some research questions might need to be answered by gathering only numerical data, others by only qualitative data. However, we recommended in Chapter 2 that, for mixed methods research, attention should be paid to the research questions such that they can *only* be answered by mixed – combined – types of data, or by adopting mixed methodologies, or by having a set of purposes that can only be addressed by mixed methods, or by taking mixed samples, or by having more than one researcher on the project (mixed researchers), in short, by building a mixed methods format into the very heart of the research. So, a research question in this vein might combine 'how' and 'what' into the same research question, or 'why' and 'who' might be combined in the same question, or description and explanation might be combined, or prediction, explanation and causation might be combined, and so on. We provide examples of these in Chapter 2.

It has been suggested (e.g. Bryman, 2007b) that, in mixed methods research, the research question has considerable prominence in guiding the research design and sampling, yet it is often more difficult to frame research questions in mixed methods research than in single paradigm research (e.g. quantitative or qualitative) (Onwuegbuzie and Leech, 2006a, p. 477). This is because it requires quantitative and qualitative matters to be addressed within the same research questions. Onwuegbuzie and Leech provide examples of mixed methods research questions, such as 'What is the relationship between graduate students' levels of reading comprehension and their perceptions of barriers that prevent them from reading empirical research articles?' (pp. 483–4). Here both numerical and qualitative data are required in order to provide a complete answer to the research question (e.g. numerical data on levels of reading comprehension, and qualitative data on barriers to reading articles) (p. 484). They provide another example of mixed methods research questions thus: 'What is the difference in perceived classroom atmosphere between male and female graduate students enrolled in a statistics course?' (p. 494). This could involve combining measures with interviews.

Here is not the place to discuss the framing of research questions (Chapter 10 addresses this). Here we draw attention to research questions per se, in particular their clarity, ease of answering, comprehensiveness, comprehensibility, specificity, concreteness, complexity, difficulty, contents, focus, purposes, kinds of data required to answer them and utility of the answers provided, to enable researchers to decide whether the particular piece of research is worth pursuing. This will require researchers to pause, generate and reflect on the kinds of research question(s) required before they decide whether to pursue a particular investigation. This argues that researchers may wish to consider whether they really wish to embark on an inquiry whose research questions are too difficult or complex to answer within the scope or time frames of the study. Many of the most useful pieces of research stem from complex issues, complex research questions and 'difficult-to-answer' research questions. They move from Alvesson's and Sandberg's (2013) 'gap filling' to problematization, new ideas and areas, innovatory thinking and the elements that make for Davis's (1971) 'interesting' research, mentioned in Chapter 4.

9.7 The literature search and review

A distinction has to be drawn between a literature search and a literature review. The former identifies the relevant literature; the latter does what it says: reviews the literature selected. If the researcher knows in advance what are the research purposes, issues and research questions then this can make the literature

search efficient, directed and selective; they determine what to look for. But this is not always the case. It is frequently the case that the researcher does not have an exact or clear picture of the field or what is relevant, and is relying on the literature review to provide such clarity and exactitude. In this situation, the literature search risks being somewhat aimless, too wide or too unfocused. In Chapter 11 we provide detailed guidance on how and where to conduct a literature search. Among other kinds of written or online materials, a sound literature search (and indeed review) will include up-to-date information from materials such as: books, articles, reports, research papers, newspaper articles, conference papers, theses, dissertations, reviews and research syntheses, government documents, databases and Internet sources, primary and secondary sources and so on.

A literature review is an essential part of many kinds of research, particularly if the research is part of a thesis or dissertation. It serves many purposes, for example:

- it ensures that the researcher's proposed research will not simply recycle existing material (reinventing the wheel), unless, of course, it is a replication study;
- it gives credibility and legitimacy to the research, showing that the researcher has 'done his/her homework' and knows the up-to-date, key issues and the theoretical, conceptual, methodological and substantive problems in the field in which the research is being proposed;
- it clarifies the key concepts, issues, terms and the meanings of these for the research;
- it acts as a springboard into the study, raising issues, showing where there are gaps in the research field, and providing a partial justification or need for the research. It makes clear where new ground has to be broken in the field and indicates where, how and why the proposed research will break that new ground;
- it indicates the researcher's own critical judgement on prior research or theoretical matters in the field and, indeed, provides new theoretical, conceptual, methodological and substantive insights and issues for research;
- it sets the context for the research and establishes key issues to be addressed;
- it enables the researcher to raise questions that still need to be answered in the field, how to move the whole field forward, and how to look differently at the field;
- it establishes and justifies the theoretical and conceptual frameworks of the research and the research design (see also Chapter 4).

We provide more details on conducting the literature search and review in Chapter 11. A literature review must be useful, not only to show that the researcher has read some relevant materials, as this is a trivial, self-indulgent reason, but that this actually informs the research. A literature review must be formative and lead into, or give rise to, all aspects of the research: the field, the particular topic, the theoretical grounding and framework, the methodology, the data analysis and implications for future research.

The researcher who is contemplating conducting a particular piece of research will need to give careful consideration to the necessary size and scope of the literature review, as this has implications for time, manageability, practicability and decision making on whether the project is too large, unfocused, diffuse, general or difficult to have justice done to it in the time and resources available. It is a determinant of whether to opt for a particular piece of research.

9.8 Summary of key issues in choosing a research topic or project

This chapter has set out several practical considerations in choosing a research topic. We advise researchers, both novice and experienced, to approach the selection of, and decision making on, a research topic with caution, going into it 'with their eyes open', aware of its possible pitfalls as well as its benefits and implications. We summarize the points discussed in the chapter in Box 9.1.

BOX 9.1 ISSUES TO BE FACED IN CHOOSING A PIECE OF RESEARCH

1. Make the topic small. Think small rather than big.
2. Limit the scope and scale of the research: think narrow rather than broad.
3. Keep the focus clear, limited, bounded and narrow.
4. Don't be over-ambitious.
5. Be realistic on what can be done in the time available, and whether, or how much, this might compromise the viability or worth of the research.
6. Make it clear what has given rise to the research – why choose this topic/project.
7. Choose a topic that might enable you to find your niche or specialism in the research or academic world or which might help to establish your reputation.
8. Decide why the research is important, topical, interesting, timely, significant, original, relevant and positively challenging.
9. Decide what contribution the research will make to the conceptual, practical, substantive, theoretical and methodological fields.
10. Decide whether your research is mainly to 'fill a gap' or to break new ground, to be innovatory.
11. Choose a research project that will be useful, and decide how and for whom it will be useful.
12. Decide why your research will be useful and who will/might be interested in it.
13. Decide what might be the impact of your research, and on whom.
14. Choose a topic that is manageable and practicable.
15. Choose a topic that will enable rigour to be exercised.
16. Choose a topic that has clear boundaries or where clear, realistic, fair boundaries can be set.
17. Decide what the research will 'deliver'.
18. What will the research do?
19. What will the research seek to find out?
20. Choose a topic for which there is a literature.
21. Decide whether you will have the required access and access to what/whom in order to be able to conduct the research.
22. Decide what can and cannot be done within the time and timescales available.
23. Decide what can and cannot be done within the personal, people-related, material, effort-related, financial and scope of the research.
24. Consider the likely clarity, scope, practicability, comprehensiveness, ease of answering, framing, focus, kinds of data required, comprehensibility of the research questions and their combination.
25. Consider whether the research will influence, or make a difference to, practice, and, if not, why it might still be important.
26. Consider whether you have the right personality, characteristics, experience and interpersonal behaviour to conduct the proposed piece of research.
27. Consider whether the research will sustain your creativity, imagination, positive attitude and motivation over time.
28. Choose a topic for which you know you will able to receive expert, informed supervision.
29. Be clear on why you – personally, professionally, career-relatedly – want to do the research, and what you personally want out of it, and whether the research will enable you to achieve this. How will the research benefit you?
30. How will the research benefit the participants?
31. How will the research benefit the world of education?
32. Choose a topic that will sustain your interest over the duration of the research.
33. Consider whether you have sufficient experience, skills and expertise in the field in which you want to conduct the research for you to be able to act in an informed way.
34. Consider whether it is advisable to embark on a piece of research that deliberately does not have research questions.
35. Consider the necessary complexity (where it exists) of the research phenomenon, scope and conduct of the research, and the difficulty of the research issues, foci and conduct.
36. Consider how future research will be able to build on your research, i.e. that the research opens up possibilities rather than closes them down.

Companion Website

The companion website to the book provides PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Research questions

CHAPTER 10

This chapter will explore:

- the purpose of research questions and where they come from
- different kinds of research questions
- devising your research question(s)
- making your research question answerable
- how many research questions you should have

10.1 Why have research questions?

Research design includes a concrete and specific statement of the aims and objectives of the research as set out in the overall research purposes. There is a move in the research design from the general to the specific and concrete. From these specific, concrete objectives the researcher can formulate direct, concrete, specific research questions that the research will answer specifically and concretely and, thereby, address the objectives of the research. Research questions get to the heart of the research issue.

For many kinds of research, the framing of the research question(s) is critical; it focuses, centres, shapes, steers and drives the entire research and it is the answers to the research questions in which the researcher is interested. As Alvesson and Sandberg (2013) remark, research questions concern the direction of a study and what it is about (p. 2). They strive to 'tame curiosity' (White, 2013, p. 213) and to shape and direct the research (Agee, 2009), to make the research topic tractable. Research questions might raise a problem and shape it into a testable question or hypothesis and enable the results to be reported; they inform the direction of the research in substantive, contextual, theoretical and methodological terms; in other words, they indicate what the research is really about and what it must address.

Research questions are not the start of the research; typically they stem from the overall research purposes, objectives and design. They are the concrete questions, carefully composed in order to address the research objectives, to constitute a fair operationalization and embodiment of a valid set of indicators for addressing the research objectives, providing answers which address the research purposes with warranted data. Research questions render research aspirations, in principle, researchable and able to be investigated scientifically and rigorously, and answered empirically or by appropriate non-empirical means. We say 'in principle' because other factors, for example, practical matters such as access, permissions, finances and resources (human, material, temporal, administrative), may obstruct the research progress. Research questions take the purposes and objectives of the research and narrow them down into specific, concrete areas of focus; they narrow the boundaries of the research and help the researcher to decide where to go in the research.

This chapter does not distinguish between qualitative and quantitative research, as the issues raised apply to both. It is invidious to suggest that certain issues apply only to quantitative research and that others apply only to qualitative research; the issues apply to both types, and, indeed, mixed methods research demonstrates this very clearly, drawing on different kinds of research and data in order to answer a particular research question. For example, Simon (2011) notes that qualitative research questions tend to be exploratory and open in nature (p. 1), but there is no reason why this cannot apply to quantitative research.

Research questions typically precede the specification of research designs, methodologies, data types, methods of data collection, instrumentation and sampling, i.e. the logistical aspects of the research and which follow from the research questions.

10.2 Where do research questions come from?

Research questions stem from the aims, purposes and objectives of the research. Research questions turn a general purpose or aim into specific questions to which specific, data-driven, concrete answers can be given. This is the process of operationalization of the aims and purposes into research questions. Researchers must ensure that there is an alignment between the aims and objectives of the research and the research questions,

such that the latter serve the former. The research questions must yield data that provide warrantable evidence to address the research purposes and objectives and to draw conclusions. They must follow logically from the research purposes and objectives, and the data used in answering them must be reliable and valid indicators of the evidence needed to answer the research purposes and objectives.

It is the answers to the research questions that can provide some of the 'deliverables' referred to in Chapter 9. A useful way of deciding whether to pursue a particular study is to ascertain the clarity and ease with which research questions can be conceived and answered. Leong *et al*. (2012) argue that, in constructing research questions, it is important to have: (i) knowledge of the literature on the topic (research literature, theoretical literature); (ii) an awareness of the implications, practicability and limitations in conducting the research; and (iii) an integration of (i) and (ii). Whereas the overall research identifies the *field*, the *main topic* and *direction* of the research, the research question asks for specific, explicit answers from the outcomes of the research (p. 34).

For example, take the issue 'why do females have higher scores than males in international tests of reading at age 14?'; here the research questions might ask: (a) 'what are the test scores of females and males in such-and-such an international test of reading comprehension at age 14 in such-and-such a country?'; (b) 'how consistent among different sub-groups of females and males are the scores in such-and-such an international test of reading comprehension at age 14 in such-and-such a country?'; (c) 'how much variation is there in the scores of females and males in the scores in such-and-such an international test of reading comprehension at age 14 in such-and-such a country?'; and (d) 'what reasons do the test designers and data give for the answers to (a), (b) and (c)?'. Here the initial single overall question generates several research questions; this is common, as one of the purposes of a 'good' research question is to take a particular objective of the research and render it concretely researchable and practicable (White, 2009, p. 34).

10.3 What kinds of research question are there?

Questions such as 'what is happening?', 'what has happened?' 'what might/will/should happen?' open up the field of research questions. Chapter 6 also mentioned causal questions; 'what are the effects of such-and-such a cause?' and 'what are the causes of such-and-such an effect?' are two such questions, to which can be added the frequently used questions 'how?' and 'why?'. These questions ask for explanations as well as reasons. De Vaus (2001, p. 1) notes that there are two fundamentals of research questions: 'what is going on?' (description) and 'why is it going on?' (explanation). These are useful pointers when starting to think about research questions.

A useful approach to framing different kinds of research questions can be to ask questions that start with: what; what if; who; when; where; which; whence; whither; why; and how. There are many categories or types of research question. An early typology of these stem from Dillon (1984) who identified seventeen types of research question, which he refined into four main types: descriptive, explanatory, comparative and normative. His 'first order' type addresses 'properties' (p. 330): existence, identification, affirmation, substance, definition, character, function and rationale. His 'second order' type concerns 'comparisons': concomitance, conjunction and disjunction, equivalence and difference. His 'third order' type concerns 'contingencies': relations, correlations, conditionality (consequence and antecedence) and causality. His 'extra order' type concerns deliberation (normative questions), and other attributes. He arranges these in a hierarchy, with causal questions at the apex, being closest to the purpose of scientific inquiry.

Flick (2009) differentiates questions concerning describing states (what they are, how they came about, how they are sustained) from those describing processes (how and why something develops or changes) (p. 102). He also distinguishes between those questions which seek to confirm existing hypotheses or assumptions and those which seek to discover or allow new assumptions or hypotheses (p. 102), the latter being Strauss's (1987) 'generative questions', which are those that 'stimulate the line of investigation in profitable directions; they lead to hypotheses, useful comparison, the collection of certain classes of data, even to general lines of attack on potentially important problems' (Strauss, 1987, p. 22).

Agee (2009, p. 433) reports four kinds of research question: exploratory, explanatory, descriptive and emancipatory. Denscombe (2009a) identifies six types, articulated with their concern: description, prediction, explanation, evaluation, development-related and empowerment. De Vaus (2001) adds 'comparison' to these. Research questions can concern, for example:

- prediction ('what if' and 'what will' types of question), understanding, exploration, explanation (reasons for: 'why-type' questions; 'how-type' questions), description ('what-type' questions) and causation;

- testing and evaluation;
- comparisons/relations/correlations (between variables, people, events);
- processes, functions and purposes; stages of something;
- factors, structures, properties and characteristics of something;
- classification, types of something, trends and patterns;
- how to achieve certain outcomes; how to do, achieve, improve and develop something; alternatives to something;
- empowerment (of individuals and groups).

White (2009, pp. 42–4) argues that 'metaphysical questions' (those which cannot be answered completely through empirical research and observation) and 'normative questions' (those concerning judgements of values, what 'should' or ought to be the case or should happen, ethical and moral matters: what is desirable, good, bad, right, wrong, defensible) are typically beyond the scope of empirical social science, being 'deliberative' questions (p. 43) to which there are multiple answers deriving from people's opinions. Similarly, Hammersley (2014) comments that such questions are out of court for social scientists. Social science, he avers, should concern itself with factual data (descriptions and explanations), and social scientists have no more authority than others to determine what is good or bad (pp. 94, 144).

10.4 Devising your research question(s)

Research questions should enable the researcher to make a significant and innovative contribution to the field of study, say something new and interesting and contribute to the concerns and current topics in the academic community (see Chapter 4). Researchers should check that their research question will yield useful, relevant and significant data on matters that recipients (widely defined) of the research will care about (the 'so what?' criterion). It is also useful to consider whether the research question is 'gap filling', 'neglect filling', a new formulation of an existing idea or an entirely new idea, and how the facts which the answers to the research yield will match relevant theory.

Researchers need to decide exactly what they need to know about the matter in hand and make sure that, together, the research questions address all the required scope of the research. Though it sounds like common sense, it is important to check that it is possible to answer the research questions and that the answers to the research questions stem from data. The research questions must be manageable, practicable and answerable, fully operationalized, with a clear delineation of their scope and boundaries, and that they can be answered within the time frame and scope of the research.

With regard to the formulation of the research questions there are several points to make:

- Make sure that the types of research questions are fit for purpose (e.g. descriptive, explanatory, causal, evaluative, exploratory etc.) and that the research questions suggest an appropriate methodology. Where relevant, ensure that your research questions will be amenable to formulating hypotheses.
- Make your research questions as brief, clear, specific, concise and precise as possible (no more than a single sentence) (White, 2009, pp. 66–70), ensuring that they address (a) the focus: the 'what'; (b) the persons: the 'who' (the population and the sample as appropriate); (c) the location (the 'where'); and (d) the timing (the 'when' or the (historical) period studied) of the research (pp. 71–2).
- If you have more than one research question, make clear the relationship (e.g. logical) between them and the relative status of each question (is one question more important than another, and, if so, why or do they have equal status?) (cf. Andrews, 2003, p. 35).
- If you have one research question with several subsidiary questions (discussed later in this chapter), make clear the relationship (such as logical, chronological, empirical) not only between the subsidiary questions but between them and the main research question. Identify the main research question and the contributing subsidiary research questions (if there are any) (cf. Andrews, 2003).
- Check whether some of your research questions are more general/specific than others, and, if so, why. Check the scope of the research question: make sure your research questions are very focused, neither too narrow nor too broad. Avoid questions that require a simply binary response (yes/no). Avoid personal pronouns in the research questions.

Lipowski (2008, p. 1669) suggests that researchers can examine the four s's of research questions in order to determine their importance: size (the magnitude of an effect); scope (the overall effect on existing practice); scalability (how the findings may have expanded – wider – impact); and sustainability (long-term effects and support). It is useful to ask a colleague to review one's research questions and to give feedback on them.

White (2009) provides some useful cautions in constructing research questions:

- Only ask one question at a time (p. 37). Avoid putting two questions into the same single question, as it is important to see which answer refers to which part of the question. For example, avoid putting into the same research question a 'what' and 'why' question; they are asking for two different kinds of response/data, for example, 'what are the test scores of females and males in such-and-such an international test of reading comprehension at age 14 in such-and-such a country and how can we account for such findings?'. Combining descriptive, explanatory, causal, comparison, correlational, evaluative or other types of question into a single research question builds in questionable ambiguity. However, as discussed in Chapter 2, mixed methods research often suggests combining more than one question in a research question.
- Avoid 'false dichotomies' (p. 37). For example, in the question 'is a country's centralized university entrance examination a narrowing of the curriculum or a fair basis for comparing student performance?', neither or both statements may be true, partially true, irrelevant, or, indeed, there may be a less polarized position.
- Avoid making false assumptions (p. 38). For example, in the question 'why do males prefer multiple choice questions to essay questions in public English language examinations at age 16?', there are suppressed assumptions that such a preference exists, that multiple-choice questions are all of a single type (and the same applies to essay questions), that English language examinations are of a single type, and so on – many questionable assumptions and ambiguities underlie the research question. Whilst it may be impossible, because language and terminology inherently carry ambiguities, to render research questions unambiguous, nevertheless the researcher should avoid making false assumptions; in other words, the assumptions made should be warrantable.
- Avoid tautological questions (p. 40), i.e. those questions which say the same thing in more than one way. For example, in the question 'why do so many wealthy students study in elite universities?', one of the criteria (among others, of course) for a university to be regarded as 'elite' is that it recruits from among the wealthy groups in society. In other words, the research question here could be rewritten as 'why do so many wealthy students study in universities which recruit mainly wealthy students?' As White (2009, p. 41) remarks, this type of question is redundant because it already supplies its answer.

One can add to these cautions:

- Avoid making the research question too broad. For example, a research question such as 'what are the effects of such-and-such an intervention on students?' is far too broad, and could be replaced by, for example: 'how does such-and-such an intervention relate to sixteen-year-olds' examination performance in mathematics?'.
- Avoid making research questions too simple. For example, 'how are schools addressing student under-achievement?' could be answered by a simple Internet search, whereas a more complex question could be 'what are the effects of such-and-such an intervention in upper primary schools on the achievement of students at age 11?'.
- Avoid biased and leading questions (Agee, 2009), avoid 'can'/'how can' questions, as these are hypothetical and limitless (Andrews, 2003, p. 34).
- Avoid making your research question your questionnaire question; the former is overall and the latter is specific (Andrews, 2003, p. 69).

Some authors set out a linear process of devising research questions (cf. Alvesson and Sandberg, 2013, pp. 21–2), for example:

Step 1: Identify the field of study/subject area.
Step 2: Identify a specific topic within the field of study.
Step 3: Identify the purpose of the particular study.
Step 4: Formulate a research question that relates to the specific topic which is of both theoretical and practical interest/concern.

Leong *et al*. (2012, p. 127) suggest an alternative sequence:

Step 1: Define the domain of the research.
Step 2: Identify the main factors in, attributes of, conceptual frameworks of, influences on, and practical implications of, the topic in question.
Step 3: Plan how to cover these main factors/attributes/influences/conceptual frameworks/implications in formulating your research question, including which ones to address or leave aside.
Step 4: Operate a convergent exercise in bringing steps (1) to (3) into a researchable question (the authors recommend mixed methods in preference to either quantitative or qualitative methods, as this is consistent with their advocacy of 'multiple and convergent operationalism').

However, Alvesson and Sandberg (2013) suggest that, in reality, the formulation of a research question is much more iterative, interactive and evolutionary than that which is set out in a simple linear approach, and includes greater reference to literature, current debates and policy concerns. Leong *et al*. (2012) advocate brainstorming ideas, from which practicable, interesting and novel research questions can be selected; this might involve connecting ideas that may not have previously been connected ('novel links') (p. 120) and trying to look at a phenomenon as an outsider might view it. In this respect, mixed methods may possess greater potential for effective research questions than mono-methods approaches (see Chapter 2).

Similarly, researchers should evaluate their research questions and be prepared to modify them either before or during the research (if appropriate). As research progresses, matters may arise which indicate that the initial research question was too broad, or that the focus needs to shift, or that a more specific question needs to be asked. Research questions can change over time, as the researcher becomes more immersed in the research and as the research unfolds over time. This is commonplace and is almost to be expected: as the research becomes more refined, so the research questions will become more refined. The point here is that, at the start of the research it is not always clear where the research will go, and this means that the research question(s) could well change over time as the phenomenon in question is unpacked.

Similarly, what the researcher initially planned or wished to do in the research may have to be modified as the actual research is negotiated or unfolds. As Chapter 13 makes clear, this is not uncommon in sensitive research, but it is not confined to that: what the researcher wishes to do and what he/she can do in reality are not the same, and this may affect the research questions. A range of practical constraints, such as time, resources, access, scope can lead to research questions being modified over time. Further, as the research unfolds, unforeseen avenues for important exploration may open up, or what the researcher had initially thought was the 'correct' research question may turn out to need modification in order to get to the heart of the matter. This, too, is not uncommon; indeed in some kinds of research (e.g. ethnographic and qualitative research) it may even be expected to occur.

Some research – often qualitative (Bryman, 2007b) – may not have research questions. Similarly, it is important to recognize that research methods are not always driven by the research questions (p. 18), and that one should avoid the 'dictatorship of the research questions' (p. 14) in steering the design and conduct of the enquiry. Nevertheless, in many kinds of research the research questions figure significantly, and hence the chapter moves to considering their importance.

Some kinds of research (e.g. ethnography) might not begin with research questions but, in their closing stages, might use the open-ended research (e.g. an ethnography, interviews, focus groups) to raise research questions for further study in subsequent investigations. Such research, being exploratory in nature, might not wish to steer the inquiry too tightly, and indeed one of the features of naturalistic research (see Chapter 15) is that it endeavours not to disturb the everyday, natural setting for the participants. However, for many kinds of research, one of the early considerations that researchers can address in choosing a project is the research questions that the study might generate (or indeed should, as they derive from the overall purposes of the research).

In considering the proposed research, a useful approach is to brainstorm the possible areas of the field, moving from a general set of purposes to a range of specific, concrete issues and areas to be addressed in the research, and, for each, to frame these in terms of one or more research questions (or indeed in terms of a thesis to be defended or a hypothesis to be tested).

10.5 Making your research question answerable

There are many different kinds of research questions that derive from different purposes of the research. For example, research questions may seek:

- to describe what a phenomenon is and what is, or was, happening in a particular situation (e.g. in ethnographies, case studies, complexity theory-based studies, surveys);
- to explain why something happened;
- to predict what will happen (e.g. in experimentation, causation studies, research syntheses);
- to investigate what should happen (e.g. in evaluative research, policy research, ideology critique, participatory research);
- to examine the effects of an intervention (e.g. in experimentation, ex post facto studies, case studies, action research, causation studies);
- to examine perceptions of what is happening (e.g. in ethnography, survey);
- to compare the effects of an intervention in different contexts (experimentation, comparative studies);
- to test a theory or hypothesis;
- to develop, implement, monitor and review an intervention (e.g. in participatory research, action research).

In all of these the task of the researcher is to turn the general purposes of the research into actual practice, to operationalize the research. We discuss the process of operationalization in Chapter 11. In the present chapter we note that operationalization in terms of research questions means moving from very general, broad questions to very specific, concrete, practicable questions to which specific answers can be given. Thus the researcher breaks down each general research purpose or general aim into more specific research purposes and constituent elements, continuing the process until specific, concrete questions have been reached to which specific answers can be provided. This is not unproblematic; for example, Leong *et al*. (2012) note that operationalization, whilst valuable, may be prone to rendering issues biased or simplistic, and that, to overcome this, it is important to consider multiple perspectives on, and methodologies for researching, the topic (triangulation) (p. 127). Two examples of operationalization are provided below.

Let us imagine that the overall research aim is to ascertain the continuity between primary and secondary education (Morrison, 1993, pp. 31–3). This is very general, and needs to be translated into more specific terms. Hence the researcher might deconstruct the term 'continuity' into several components, for example, experiences, syllabus content, teaching and learning styles, skills, concepts, organizational arrangements, aims and objectives, ethos, assessment. Given the vast scope of this, the decision is taken to focus on continuity of pedagogy. This is then broken down into its component areas: the level of continuity of pedagogy; the nature of continuity of pedagogy; the degree of success of continuity of pedagogy; the responsibility for continuity; record-keeping and documentation of continuity; resources available to support continuity.

The researcher might take this further into investigating: the *nature* of the continuity (the provision of information about continuity); the *degree* of continuity (a measure against a given criterion); the *level of success* of the continuity (a judgement). An operationalized set of research questions, then, might be:

- How much continuity of pedagogy is occurring across the transition stages in each curriculum area? What kind of evidence is required to answer this question? On what criteria will the level of continuity be decided?
- What pedagogical strategies operate in each curriculum area? What are the most frequent and most preferred? What is the balance of pedagogical strategies? How is pedagogy influenced by resources? To what extent is continuity planned and recorded? On what criteria will the nature of continuity be decided? What kind of evidence is required to answer this question?
- On what aspects of pedagogy does planning take place? By what criteria will the level of success of continuity be judged? Over how many students/teachers/curriculum areas will the incidence of continuity have to occur for it to be judged successful? What kind of evidence is required to answer this question?
- Is continuity occurring by accident or design? How will the extent of planned and unplanned continuity be gauged? What kind of evidence is required to answer this question?
- Who has responsibility for continuity at the transition points? What is being undertaken by these people?
- How are records kept on continuity in the schools? Who keeps these records? What is recorded? How frequently are the records updated and reviewed? What kind of evidence is required to answer this question?
- What resources are there to support continuity at the point of transition? How adequate are these resources? What kind of evidence is required to answer this question?

It can be seen that these questions, several in number, have moved the research from simply an expression of interest (or a general aim) into a series of issues that lend themselves to being investigated in concrete terms. This is precisely what we mean by *operationalization*. The questions above also deliberately avoid the precision that one might be seeking in some research questions, such as the delineation of the locale of the research and the schools in question.

It is now possible to identify not only the specific questions to be posed, but also the instruments that might be needed to acquire data to answer them (e.g. semi-structured interviews, rating scales on questionnaires, or documentary analysis). By operationalization we thus make a general purpose amenable to investigation, be it by measurement or some other means. The number of operationalized research questions is large here, and may have to be reduced to maybe four or five at most, in order to render the research manageable.

Take another example of operationalizing a research question: 'do students work better in quiet rather than noisy conditions?' Here it is important to define who are the 'students', what is meant by 'work better', 'quiet' and 'noisy'. 'Students' might be fifteen-year-old males and females in school, 'work better' might mean 'obtain a higher score on such-and-such a mathematics test', 'quiet' might mean 'silence', and 'noisy' might mean 'having moderately loud music playing'. Hence the fully operationalized research questions might be 'do fifteen-year-old male and female students in school obtain a

higher score on such-and-such a mathematics test when tested when there is silence rather than when there is moderately loud music playing?' Now we have defined – and thereby narrowed – the scope, terms, field, focus, location, participants, indicators (a measurable score) and the conditions (silence and moderately loud music).

In this example the process of operationalization is to break down the constructs (or abstract terms) in question into component variables (categorical, continuous, dependent and independent), which, as the term suggests, can vary, and which are describable, observable and, in this case, measurable.

Hypotheses

An alternative way of operationalizing research questions takes the form of hypothesis raising and hypothesis testing. A 'good' hypothesis has several features:

- It is clear on whether it is directional or non-directional: a directional hypothesis states the kind or direction of difference or relationship between two conditions or two groups of participants (e.g. students' performance increases when they are intrinsically motivated). A non-directional hypothesis simply predicts that there will be a difference or relationship between two conditions or two groups of participants (e.g. there is a difference in students' performance according to their level of intrinsic motivation), without stating whether the difference, for example, is an increase or a decrease. (For statistical purposes, a directional hypothesis requires a one-tailed test whereas a non-directional hypothesis uses a two-tailed test; see Part 5.) Directional hypotheses are often used when past research, predictions or theory suggest that the findings may go in a particular direction, whereas non-directional hypotheses are used when past research or theory is unclear or contradictory or where prediction is not possible, i.e. where the results are more open-ended.
- It is written in a testable form, that is, in a way that makes it clear how the researcher will design an experiment or survey to test the hypothesis (e.g. 'fifteen-year-old male and female students in school obtain a higher score on such-and-such a mathematics test when tested when there is silence rather than when there is moderately loud music playing'). The concept of interference by noise has been operationalized in order to produce a testable hypothesis.
- It is written in a form that can yield measurable results.

Here it is a small step from a research question to a research hypothesis. Both specify and manipulate variables. In the example above, converting the research question into a hypothesis leads to the following hypothesis: *people work better in quiet rather than noisy conditions*. The fully operationalized hypothesis might be *fifteen-year-olds obtain a higher score on a mathematics test when tested when there is silence rather than when there is music playing.* One can see here that the score is measurable and that there is zero noise (a measure of the noise level).

In conducting research using hypotheses, one has to be prepared to use several hypotheses (Muijs, 2004, p. 16) in order to catch the complexity of the phenomenon being researched, and not least because mediating variables have to be included in the research. For example, the degree of 'willing cooperation' (dependent variable) in an organization's staff is influenced by 'professional leadership' (independent variable) and the 'personal leadership qualities of the leader' (mediating variable) which needs to be operationalized specifically.

There is also the need to consider the null hypothesis and the alternative hypothesis (discussed in Part 5) in research that is cast into a hypothesis testing model. The *null hypothesis* states that, for example, there is *no* relationship between two variables, or that there has been *no* difference in participants' scores on a pre-test and a post-test of history, or that there is *no* difference between males and females in respect of their science examination results. The *alternative hypothesis* states, for example: there *is* a correlation between motivation and performance; there *is* a difference between males' and females' scores on science; there *is* a difference between the pre-test and post-test scores on history. The alternative hypothesis is often supported when the null hypothesis is 'not supported': if the null hypothesis is not supported then the alternative hypothesis is. The two kinds of hypothesis are usually written thus:

H_0: the null hypothesis
H_1: the alternative hypothesis

We address hypothesis-testing fully in Part 5, particularly Chapters 38 and 39.

Contrary to statements that hypotheses are the province of only quantitative methods, we hold that hypotheses can be developed and tested in both quantitative and qualitative research; we see no reason why not. Nor do we concur with the view that a 'variable' is not a property of qualitative research. Theories and hypotheses can be tested in both qualitative and quantitative research, singly and together, and variables can comfortably be found and explored in both types (cf. White, 2013, p. 231). There is no exclusivity.

10.6 How many research questions should I have?

Whilst there are no hard and fast rules, a general principle is to have as few as necessary, but no fewer. Some researchers suggest having only one central research question with or without several subsidiaries (e.g. Andrews, 2003; Simon, 2011; Creswell, 2012). Others suggest no more than three or four (e.g. White, 2009); Creswell (2012) also suggests five to seven in qualitative research, whilst yet others (e.g. Miles and Huberman, 1994) extend this into double figures.

Andrews (2003) is very clear that there should be only one main research question, though a main research question may require 'subsidiary questions' (which are more specific and contribute to the answer to the main research question; p. 26) and 'ancillary questions' (which may not answer the main research question but which may be a consequence of, lead on from or broaden out the main research question; p. 81). Subsidiary questions, he avers (p. 43), are those that are '*on the way*' (his italics) to answering the main research question, whilst ancillary questions (those that provide useful but not strictly necessary material to answer the main research question) flow from, rather than contribute to, the main research question (p. 81). He cautions against having more than one main research question and, indeed, against having too many subsidiary questions, as these risk making the study too broad or ambitious in scope.

Whether one has several research questions or one research question with one or more subsidiary questions, Andrews (2003, p. 80) makes the important point that it is essential to establish the relationship (e.g. logical, chronological) between them and to identify which are the main questions and which questions are closely related or more distantly related to each other (p. 80), and how and why. His suggestion of having only one main research question is useful in identifying and focusing on the key purpose of the research.

Answering 'how many research questions do I need?' concerns the purposes of the research, the research design, the scope and magnitude of the research and each research question (and, where relevant, its subsidiaries and ancillaries) and, hence, its manageability. If the researcher wishes to avoid Andrews' suggestion of only a single, main research question, a general guide might be to have no more than four main research questions (though some would suggest that this is too many) with only two or three subsidiaries for each, but this is highly contestable and others would argue for fewer. If you have too many research questions then this might indicate that the scope of the research is too broad and ambitious, is impractical, lacks focus, lacks precision and specificity, is poorly operationalized and is insufficiently thought through. In our experience, many novice researchers have maybe three research questions, but this is very fluid.

Many studies may have one research question that asks for descriptive data, together with another that asks for explanations (causal – why – or 'how' questions), together with a third that asks for the implications/recommendations that derive from the answers to the preceding two research questions, moving from description to analysis/explanation to evaluation/implications/recommendations, i.e. three research questions (cf. Gorard, 2013, p. 37). Or the research questions may comprise: (i) a question that asks for descriptive data (what, who, where, when); followed by (ii) a question that requires comparisons, differences, relations to be drawn; followed by (iii) a question that asks 'so what?' (implications and recommendations).

10.7 A final thought

Researchers may wish to ponder on whether they want to embark on investigations that have no clearly defined research questions (cf. Andrews, 2003, p. 71) or indeed any research questions, for example an ethnography, a naturalistic observational study, studies in the humanities and arts (p. 71), or qualitative research (Bryman, 2007b). A research question may lead to a subsequent hypothesis, but that is an open question.

Companion Website

The companion website to the book provides PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Research design and planning

CHAPTER 11

This chapter sets out a range of key issues in planning research, including:

- research design and methodology
- approaching research planning
- a framework for planning research
- conducting and reporting a literature review
- searching for literature on the Internet
- how to operationalize research questions
- data analysis
- presenting and reporting the results
- a planning matrix for research
- managing the planning of research
- ensuring quality in the planning of research

It also provides an extended worked example of planning a piece of research.

Research design has to take account of the context of research and constraints on it, as these will inform orienting decisions. Such decisions are strategic; they set the general nature of the research. Here, questions that researchers may need to consider are:

- Who wants the research?
- Who will receive the research/who is it for?
- Who are the possible/likely audiences of the research?
- What powers do the recipients of the research have?
- What are the general aims and purposes of the research?
- What are the main priorities for and constraints on the research?
- Is access realistic?
- What are the timescales and time frames of the research?
- Who will own the research?
- At what point will the ownership of the research pass from the participants to the researcher and from the researcher to the recipients of the research?
- Who owns the data?
- What ethical issues are to be faced in undertaking the research?
- What resources (e.g. physical, material, temporal, human, administrative) are required for the research?

Decisions here establish some key parameters of the research, including some political decisions (e.g. on ownership and on the power of the recipients to take action on the basis of the research). At this stage the overall feasibility of the research will be addressed.

11.1 Introduction

A research design is a plan or strategy that is drawn up for organizing the research and making it practicable, so that research questions can be answered based on evidence and warrants. Some researchers argue that a research design should go into considerable detail on data-collection instruments and data types; others argue that this is a logistical rather than a logical matter, and that a design comprises only, or mainly, a logical argument in which all the elements of the argument cohere (e.g. issues of research questions, methodologies/kinds of research suitable to answer the research questions). As Labaree (2013) remarks, the research design

> refers to the overall strategy that you choose to integrate the different components of the study in a coherent and logical way, thereby, ensuring you will effectively address the research problem; it constitutes the blueprint for the collection, measurement, and analysis of data.
>
> (p. 1)

There is no single blueprint for planning research. Research design is governed by 'fitness for purpose'. The purposes of the research determine the design of the research which, in turn, informs the methodology. For example, if the purpose of the research is to map the field, or to make generalizable comments, then a survey design might be desirable, using some form of stratified sample; if the effects of a specific intervention are to be evaluated then an experimental or action research design may be appropriate; if an in-depth study of a particular situation or group is important then an ethnographic design might be suitable.

It is possible to identify a set of issues in design that researchers need to address, regardless of the specifics

of their research. This chapter indicates those matters which need to be addressed in practice so that an area of research interest can become practicable and feasible. The chapter indicates how research can be operationalized, i.e. how a general set of research aims and purposes can be translated into a practical, researchable topic.

It is essential to try as far as possible to plan every stage of the research. To change the 'rules of the game' in midstream once the research has commenced is a sure recipe for problems, though sometimes matters arise which necessitate this. The terms of the research and the mechanism of its operation must be ironed out in advance as far as possible if it is to be credible, legitimate and practicable. Once they have been decided, the researcher is in a positive position to undertake the research. The setting up of the research is a balancing act, for it requires the harmonizing of *planned possibilities* with *workable, coherent practice*, i.e. the resolution of the difference between what could be done/what one would like to do and what will actually work/what one can actually do, for, at the end of the day, research has to work. In planning research there are two phases – a divergent phase and a convergent phase. The divergent phase will open up a range of possible options facing the researcher, whilst the convergent phase will sift through these possibilities, see which ones are desirable, which ones are compatible with each other, which ones will actually work in the situation, and move towards an action plan that can realistically operate. This can be approached through the establishment of a framework of planning issues.

11.2 Approaching research planning

What the researcher does depends on what the researcher wants to know and how she or he will go about finding out about the phenomenon in question. The planning of research depends on the kind(s) of questions being asked or investigated. This is not a mechanistic exercise, but depends on the researcher's careful consideration of the purpose of the research (see Chapter 10) and the phenomenon being investigated (see Table 11.1).

Chapters 1 and 2 set out a range of paradigms which inform and underpin the planning and conduct of research, for example:

- positivist, post-positivist, quantitative, scientific and hypothesis-testing
- qualitative
- interpretive, naturalistic, phenomenological and existential, interactionist and ethnographic, qualitative
- experimental
- ideology critical
- participatory
- feminist
- political
- evaluative
- mixed methods.

It was argued that these paradigms rest on different ontologies (e.g. different views of the essential nature or characteristics of the phenomenon in question) and different epistemologies (e.g. theories of the nature of knowledge, its structure and organization, and how we investigate knowledge and phenomena: how we know, what constitutes valid knowledge, our cognition of a phenomenon). For example:

- a positivist paradigm rests, in part, on an objectivist ontology and a scientific, empirical, hypothesis-testing epistemology;
- a post-positivist paradigm rests on the belief that human knowledge is conjectural, probabilistic, influenced by the researcher and the theoretical lenses being used (i.e. there are no absolute truths or value-free enquiry), and that the warrants used to

TABLE 11.1 PURPOSES AND KINDS OF RESEARCH

Kinds of research purpose	*Kinds of research*
Does the research want to test a hypothesis or theory?	Experiment, survey, action research, case study
Does the research want to develop a theory?	Ethnography, qualitative research, grounded theory
Does the research need to measure?	Survey, experiment
Does the research want to understand a situation?	Ethnographic and interpretive/qualitative approaches
Does the research want to see what happens if…?	Experiment, participatory research, action research
Does the research want to find out 'what' and 'why'?	Mixed methods research
Does the research want to find out what happened in the past?	Historical research

support conjectures are mutable. Like positivism, it holds to a realist ontology and, unlike positivism, it holds to a conjectural, falsificationist epistemology;

- an interpretive paradigm rests, in part, on a subjectivist, interactionist, socially constructed ontology and on an epistemology that recognized multiple realities, agentic behaviours and the importance of understanding a situation through the eyes of the participants;
- a paradigm of ideology critique rests, in part, on an ontology of phenomena as organized both within, and as outcomes of, power relations and asymmetries of power, inequality and empowerment, and on an epistemology that is explicitly political, critiquing the ideological underpinnings of phenomena that perpetuate inequality and asymmetries of power to the advantage of some and the disadvantage of others, and the need to combine critique with participatory action for change to bring about greater social justice;
- a mixed methods paradigm rests on an ontology that recognizes that phenomena are complex to the extent that single methods approaches might result in partial, selective and incomplete understanding, and on an epistemology that requires pragmatic combinations of methods – in sequence, in parallel, or in synthesis – in order to fully embrace and comprehend the phenomenon and to do justice to its several facets.

Researchers need to consider not only the nature of the phenomenon under study, but also what are or are not the ontological premises that underpin it, the epistemological bases for investigating it and conducting the research into it. These are points of reflection and decision, turning the planning of research from being solely a mechanistic or practical exercise into a reflection on the nature of knowledge and the nature of being.

On the other hand some researchers argue against the need for the articulation of research paradigms in conducting research. For example, Gorard (2012) remarks:

> [i]in buying a house we would not start with epistemology, and we would not cite an 'isms' or Grand Theory. Nor would we need to consider the 'paradigm' in which we were working.… We would collect all available and any evidence available to us as time and resources allow, and then synthesize it quite naturally and without considering mixed methods as such.
>
> (p. 6)

Having a paradigm as a whole approach to research is, for him, simply a 'red herring' (p. 7); this is contestable.

11.3 Research design and methodology

Having a rigorous research design is crucial in the research process. In planning research, the researcher commences with the overall purposes of the research and then constructs a research design to address these. De Vaus (1999) contends that a research design functions to ensure that the evidence that research obtains enables them to 'answer the initial question as unambiguously as possible' (p. 9) and to indicate the kind of evidence required to answer the research questions.

Research design is, as White (2013, p. 221) notes, a logical rather than a logistical matter, i.e. concerned with the overall blueprint – the architecture – rather than the 'nuts and bolts' of how to carry out that plan (the implementation of the plan and the building materials to be used). The 'logic' here is the sequence which connects the data (typically empirical data) to the research questions and its conclusions (Yin, 2009, p. 26). It ensures that evidence is linked to research questions and conclusions and it makes clear the logic which connects the data to the evidence.

The research design identifies the evidence needed to address the research purposes, objectives and questions, i.e. the logic that underpins the connections between purposes, objectives, questions, data and conclusions drawn. Evidence requires an indication of the warrants that will be used to support the case made from the findings of the research. In other words, the research design connects the idea and the conclusions with the evidence; it sets out the 'chain of reasoning' and the warrants that link together these elements (White, 2009, p. 112). A claim about, or conclusion from, the research needs not only an evidence base but also a logical warrant that renders the evidence a fair defence of the claim or conclusion. A warrant, then, provides the link, the 'backing' between the evidence and the proposition under study (Andrews, 2003, p. 30). Imagine a court of law: a case is made for such-and-such, and the evidence is brought to support that case. The evidence is a defensible selection of the data available.

A research design includes research questions and the nature of, and warrants for, the evidence required to answer those questions. Research design does not dictate the kinds of *data* (de Vaus, 1999, p. 9), but it indicates the kinds of *evidence* (see also Gorard, 2013, p. 6). Research design precedes decisions on data types.

Evidence is not the same as data. Data are neutral, an unsorted collection of any information or facts. Evidence is what you derive from those data, i.e. once selected, processed, organized and brought into the service of supporting a claim, argument, interpretation, proof, theory, conclusion or answer to a question, then data become evidence. Data require a warrant in order to become evidence. A warrant is

> an argument leading from the evidence to the conclusion.… [It is] the form in which people furnish rationales as to why a certain voice … is to be granted superiority … on the grounds of specified criteria.… The warrant of an argument can be considered to be its general principle – an assumption that links the evidence to the claim made from it.
>
> (Gorard, 2002, p. 137)

Data/Information + Warrant (criteria for an evidential relationship) → Evidence.

Data are just facts, states of affairs, or propositions expressing facts; data become evidence once they enter into evidential relationships; and evidential relationships are typified by *prediction*, *confirmation/refutation* and *explanation*. Suppose we have our hypothesis H, and then there are many data/propositions available; let us call them D1, D2, D3, D4 etc. Data D3 will enter into an evidential relationship with H (will be 'evidential' with respect to H), which, if true, would: *predict* that D3 would occur; be supported (*confirmed*) or *disconfirmed* (refuted, falsified) by D3; *explain* why D3 occurs (cf. Mayo, 2004, p. 79). Data are evidential by a theoretical connection made between the hypothesis and data, and this theoretical connection 'warrants' the data; it gives the data this particular kind of normative power termed 'evidential'. Hence theory is important.

An example of using a warrant might be as follows, simplified for ease of understanding. Imagine that a research study focuses on male and female student performance in the upper end of secondary schools, and finds that upper secondary school males outperform females in mathematics. The researcher concludes that teachers are responsible for the differential mathematics achievement of upper secondary school males and females. How are the data connected to the conclusion drawn? What is the warrant linking the evidence to the conclusion, and how sound is the warrant?

The data are, for example, examination results, classroom observations and interviews. The warrant here might be that teachers operate a self-fulfilling prophecy in their differential expectations of males and females and that this self-fulfilling prophecy is the major factor responsible for the differential mathematics performance. But other warrants/acceptably justified and defensible explanations are also possible, for example: (a) student motivation exerts a major influence on mathematics performance; (b) teachers' pedagogical strategies exert a major influence on mathematics performance; (c) home conditions for study exert a major influence on mathematics performance; (d) parental influence tracks males and females into different subject preferences; (e) students' intended careers track/steer males and females into according differential significance to mathematics and so on. The list of possible warrants/defensible explanations is endless, and so it is incumbent on the researcher to demonstrate that the warrant chosen – the operation of the self-fulfilling prophecy and teacher expectation – trumps the other rival warrants. Applying the logic of the present warrant will need to show that it pulls its weight in offering a more defensible explanation than other warrants. In turn, this may require additional data and evidence not only to support the warrant given but to demonstrate that rival warrants (e.g. (a) to (e) above) are not supported, or are less well supported, by relevant evidence. Gorard (2002) provides useful examples of faulty warrants in published research.

A research design will include items such as:

- the research purposes;
- the research questions;
- the problem, issue, phenomenon, matter to be addressed and the focus of the research;
- the kind of research to be undertaken (methodology(ies)), for example, longitudinal, experimental, action research, survey, ethnographic, case study, mixed methods, together with a justification for the kind chosen;
- the timing and duration of the research;
- the content of the research (which may lie on a continuum from interventionist to non-interventionist);
- the people, groups/sub-groups or cases involved and how these are decided;
- how to ensure reliability and validity in the kinds of evidence needed to meet the requirements of the warrants required (i.e. why should we believe that the answers given to the research questions provide us with fair evidence or conclusions; how convincing are the answers; how does the evidence, the findings of the research, lead to the conclusions drawn, and how safe is this, e.g. in comparison to possible alternative conclusions and interpretations);
- addressing the ethical issues in the research;
- the organization of the research.

Creswell (2012) adds to these elements of research design the data collection, analysis and reporting procedures to be used (p. 20), though this implies that the design will move beyond statements of evidence to statements of data types and instrumentation (see also Wellington, 2015), i.e. it moves towards logistical as well as logical matters. Similarly, Ragin (1994a, p. 191) and Flick (2009) note that a research design includes fine detail that ranges from data collection to techniques of data analysis.

There appears to be little consensus on the level of detail or scope of what to include in the research design, particularly in respect of whether it should include instrumentation for data collection, data types and methods of data analysis. Whether a design should include logistical rather than simply logical matters is an open question; there are powerful arguments to support and counter both views (cf. Gorard, 2013).

There are many different kinds of design, and we introduce several of these in this book, for example: experimental, survey, ethnographic, action research, case study, longitudinal, cross-section, causal, correlational. None of these indicate data types, and indeed each or all of these might use questionnaires, observational data, interviews, documents, tests, accounts and so on.

11.4 From design to operational planning

If the preceding comments are strategic then decisions in this field are tactical; they establish the practicalities of the research, assuming that, generally, it is feasible (i.e. that the orienting decisions have been taken). Decisions here include addressing such questions as:

- What are the specific purposes of the research?
- Does the research need research questions?
- How are the general research purposes and aims operationalized into specific research questions?
- What are the specific research questions?
- What needs to be the focus of the research in order to answer the research questions?
- What is the main methodology of the research (e.g. a quantitative survey, qualitative research, an ethnographic study, an experiment, a case study, a piece of action research etc.)?
- Does the research need mixed methods, and if so, is the mixed methods research a parallel, sequential, combined or hierarchical approach?
- Are mixed methods research questions formulated where appropriate?
- How will validity and reliability be addressed?
- What kinds of data are required?
- From whom will data be acquired (i.e. sampling)?
- Where else will data be available (e.g. documentary sources)?
- How will the data be gathered (i.e. instrumentation)?
- Who will undertake the research?

11.5 A framework for planning research

Planning research depends on the design of the research which, in turn, depends on: (a) the kind of questions being asked or investigated; (b) the purposes of the research; (c) the research principles informing how one is working, and the philosophies, ontologies and epistemologies which underpin them. Planning research is not an arbitrary matter. There will be different designs for different types of research, and we give three examples here.

For example, a piece of quantitative research that seeks to test a hypothesis could proceed thus:

> Literature review → generate and formulate the hypothesis/the theory to be tested/the research questions to be addressed → design the research to test the hypothesis/theory (e.g. an experiment a survey) → conduct the research → analyse results → consider alternative explanations for the findings → report whether the hypothesis/theory is supported or not supported, and/or answer the research questions → consider the generalizability of the findings.

A qualitative or ethnographic piece of research could have a different sequence, for example:

> Identify the topic/group/phenomenon in which you are interested → literature review → design the research questions and the research and data collection → locate the fields of study and your role in the research and the situation → locate informants, gatekeepers, sources of information → develop working relations with the participants → conduct the research and the data collection simultaneously → conduct the data analysis either simultaneously, on an ongoing basis as the situation emerges and evolves, or conduct the data analysis subsequent to the research → report the results and the grounded theory or answers to the research questions that emerge from the research → generate a hypothesis for further research or testing.

One can see in the examples that for one method, the hypothesis drives the research, whilst for another the

hypothesis (if, in fact, there is one) emerges from the research, at the end of the study (some qualitative research does not proceed to this hypothesis-raising stage).

A mixed methods research might proceed thus:

> Identify the problem or issue that you wish to investigate → identify your research questions → identify the several kinds of data and the methods for collecting them which, together and/or separately, will yield answers to the research questions → plan the mixed methods design (e.g. parallel mixed design, fully integrated mixed design, sequential mixed design) (see Chapter 2) → conduct the research → analyse results → consider alternative explanations for the findings → answer the research questions → report the results.

These three examples proceed in a linear sequence; this is beguilingly deceptive, for rarely is such linearity so clear. The reality is that:

- different areas of the research design influence each other;
- research designs, particularly in qualitative, naturalistic and ethnographic research, change, evolve and emerge over time rather than being a 'once-and-for-all' plan that is decided and finalized at the outset of the research;
- ethnographic and qualitative research starts with a very loose set of purposes and research questions, indeed there may not be any;
- research does not always go to plan, so designs change.

In recognition of this, Maxwell (2005, pp. 5–6) develops an interactive (rather than linear) model of research design (for qualitative research), in which key areas are mutually informing and shape each other. The five main areas of his model are:

1. *Goals* (informed by perceived problems, personal goals, participant concerns, funding and funder goals, and ethical standards);
2. *Conceptual framework* (informed by personal experience, existing theory and prior research, exploratory and pilot research, thought experiments and preliminary data and conclusions);
3. *Research questions* (informed by participant concerns, funding and funder goals, ethical standards, the research paradigm);
4. *Methods* (informed by the research paradigm, researcher skills and preferred style of research, the research setting, ethical standards, funding and funder goals, and participant concerns); and
5. *Validity* (informed by the research paradigm, preliminary data and conclusions, thought experiments, exploratory and pilot research, and existing theory and prior research).

At the heart of Maxwell's model lie the research questions (3), but these are heavily informed by the four other areas. Further, he attributes strong connections between goals (1) and conceptual frameworks (2), and between methods (4) and validity (5). The links between conceptual frameworks (2) and validity (5) are less strong, as are the links between goals (1) and methods (4). His model is iterative and recursive over time; the research design emerges from the interplay of these elements and as the research unfolds.

Though the set of issues that constitute a framework for planning research will need to be interpreted differently for different styles of research, nevertheless it is useful to indicate what those issues might be. These are outlined in Box 11.1.

BOX 11.1 THE ELEMENTS OF RESEARCH DESIGN

The elements of research design

1. A clear statement of the problem/need that has given rise to the research;
2. A clear grounding in literature for construct and content validity: theoretically, substantively, conceptually, methodologically;
3. Constraints on the research (e.g. access, time, people, politics);
4. The general aims and purposes of the research;
5. The intended outcomes of the research: what the research will do and what the 'deliverable' outcomes are;
6. Reflecting on the nature of the phenomena to be investigated, and how to address their ontological and epistemological natures;
7. How to operationalize research aims and purposes;

8 Generating research questions (where appropriate) (specific, concrete questions to which concrete answers can be given) and hypotheses (if appropriate);
9 Statements of the warrants for the research (the rationale that links evidence and conclusions);
10 The foci of the research;
11 Identifying and setting in order the priorities for the research;
12 Approaching the research design;
13 Focusing the research;
14 Research methodology (approaches and research styles, e.g.: survey; experimental; ethnographic/naturalistic; longitudinal; cross-sectional; historical; correlational; *ex post facto*);
15 Ethical issues and ownership of the research (e.g. informed consent; overt and covert research; anonymity; confidentiality; non-traceability; non-maleficence; beneficence; right to refuse/withdraw; respondent validation; research subjects; social responsibility; honesty and deception);
16 Politics of the research: who is the researcher; researching one's own institution; power and interests; advantage; insider and outsider research;
17 Audiences of the research;
18 Instrumentation, e.g.: questionnaires; interviews; observation; tests; field notes; accounts; documents; personal constructs; role-play;
19 Sampling: size/access/representativeness; type – probability: random, systematic, stratified, cluster, stage, multi-phase; non-probability: convenience, quota, purposive, dimensional, snowball;
20 Piloting: technical matters: clarity, layout and appearance, timing, length, threat, ease/difficulty, intrusiveness; questions: validity, elimination of ambiguities, types of questions (e.g. multiple choice, open-ended, closed), response categories, identifying redundancies; pre-piloting: generating categories, grouping and classification;
21 Time frames and sequence (what will happen, when and with whom);
22 Resources required;
23 Reliability and validity:
validity: construct; content; concurrent; face; ecological; internal; external;
reliability: consistency (replicability); equivalence (inter-rater, equivalent forms); predictability; precision; accuracy; honesty; authenticity; richness; dependability; depth; overcoming Hawthorne and halo effects;
triangulation: time; space; theoretical; investigator; instruments;
24 Data analysis;
25 Verifying and validating the data;
26 Reporting and writing up the research.

A possible sequence of consideration is:

Preparatory issues	→	*Methodology*	→	*Sampling and instrumentation*	→	*Piloting*	→	*Timing and sequencing*
Ontology, epistemology, constraints, purposes, foci, ethics, research question, politics, literature review	→	Approaches Reliability and validity	→	Reliability and validity Pre-piloting	→		→	

Clearly this need not be the actual sequence; for example, it may be necessary to consider access to a possible sample at the very outset of the research.

These issues can be arranged into four main areas:

1 orienting decisions;
2 research design and methodology;
3 data analysis;
4 presenting and reporting the results.

These are discussed later in this chapter. Orienting decisions are those decisions which set the boundaries or the constraints on the research. For example, let us say that the overriding condition of the research is that it has to be completed within six months; this will exert an influence on the enterprise. On the one hand it will 'focus the mind', requiring priorities to be settled and data to be provided in a relatively short time. On the other hand it may reduce the variety of possibilities

available to the researcher. Hence questions of timescale will affect:

- the research questions which might be answered feasibly and fairly (e.g. some research questions might require a long data-collection period);
- the number of data-collection instruments used (e.g. there might be enough time for only a few instruments to be used);
- the sources (people) to whom the researcher might go (e.g. there might be enough time to interview only a handful of people);
- the number of foci which can be covered in the time (e.g. for some foci it will take a long time to gather relevant data);
- the size and nature of the reporting (there might be time to produce only one interim report).

By clarifying the timescale a valuable note of realism is injected into the research, which enables questions of practicability to be answered.

Let us take another example. Suppose the overriding feature of the research is that the costs in terms of time, people and materials for carrying it out must be negligible. This, too, will exert an effect on the research. On the one hand it will inject a sense of realism into proposals, identifying what is and what is not manageable. On the other hand it will reduce, again, the variety of possibilities which are available to the researcher. Questions of cost will affect:

- the research questions which might be feasibly and fairly answered (e.g. some research questions might require: (a) interviewing, which is costly in time both to administer and to transcribe; (b) expensive commercially produced data-collection instruments, e.g. tests, and costly computer services, which may include purchasing software);
- the number of data-collection instruments used (e.g. some data-collection instruments, such as postal questionnaires, are costly for reprographics and postage);
- the people to whom the researcher might go (e.g. if teachers are to be released from teaching in order to be interviewed then cover for their teaching may need to be found);
- the number of foci which can be covered in the time (e.g. in uncovering relevant data, some foci might be costly in researcher's time);
- the size and nature of the reporting (e.g. the number of written reports produced, the costs of convening meetings).

Certain timescales permit certain types of research, for example, a short timescale permits answers to short-term issues, whilst long-term or large questions might require a long-term data-collection period to cover a range of foci. Costs in terms of time, resources and people might affect the choice of data-collection instruments. Time and cost will require the researcher to determine, for example, what will be the minimum representative sample of teachers or students in a school, as interviews are time-consuming and questionnaires are expensive to produce. These are only two examples of the real constraints on the research which must be addressed. Planning the research early on will enable the researcher to identify the boundaries within which the research must operate and what are the constraints on it.

Further, some research may be 'front-loaded' whilst other kinds are 'end-loaded'. 'Front-loaded' research is that which takes a considerable time to set up, for example to develop, pilot and test instruments for data collection, but then the data are quick to process and analyse. Quantitative research is often of this type (e.g. survey approaches) as it involves identifying the items for inclusion on the questionnaire, writing and piloting the questionnaire, and making the final adjustments. By contrast, 'end-loaded' research is that which may not take too long to set up and begin, but then the data collection and analysis may take a much longer time. Qualitative research is often of this type (e.g. ethnographic research), as a researcher may not have specific research questions in mind but may wish to enter a situation, group or community and only then discover – as they emerge over time – the key dynamics, features, characteristics and issues in the group (e.g. Turnbull's (1972) notorious study of the descent into inhumanity of the Ik tribe in their quest for daily survival as *The Mountain People*). Alternatively, a qualitative researcher may have a research question in mind but an answer to this may require a prolonged ethnography of a group (e.g. Willis's (1977) celebrated study of 'how working class kids get working class jobs, and others let them'). Between these two types – 'front-loaded' and 'end-loaded' – are many varieties of research that may take different periods of time to set up, conduct, analyse data and report the results. For example, a mixed methods research project may have several stages (see Table 11.2).

In example one in Table 11.2, in the first two stages of the research, the mixed methods run in sequence (qualitative then quantitative), and are only integrated in the final stage. In example two, in the first two stages the quantitative and qualitative stages run in parallel, i.e. they are separate from each other, and they only combine in the final stage of the research. In example

TABLE 11.2 THREE EXAMPLES OF PLANNING FOR TIME FRAMES FOR DATA COLLECTION IN MIXED METHODS RESEARCH

Example one	*Example two*	*Example three*
Qualitative data to answer research questions in total or in part, or to develop items for quantitative instruments (e.g. a numerical questionnaire survey)	Quantitative data and qualitative data in parallel to answer research questions in total or in part, or to identify participants for qualitative study	Quantitative and qualitative data together to answer research questions in total or in part and to raise further research questions
↓	↓	↓
Quantitative data to answer research questions in total or in part, or to identify participants for qualitative study (e.g. interviews)	Quantitative and qualitative data in parallel to answer research questions in total or in part	Quantitative and qualitative data to answer research questions in total or in part
↓	↓	↓
Quantitative and qualitative data to answer one or more research questions	Quantitative and qualitative data to answer one or more research questions	Quantitative and qualitative data to answer research questions in total or in part

three, the mixed methods are synthesized – combined – from the very start of the research.

The researcher must look at the timescales that are both required and available for planning and conducting the different stages of the research project.

Let us take another important set of questions: is the research feasible? Can it actually be done? Will the researchers have the necessary access to the schools, institutions and people? These issues were explored in the previous chapter. This issue becomes a major feature if the research is in any way sensitive (see Chapters 5 and 13).

11.6 Conducting and reporting a literature review

Before one can progress very far in planning research it is important to ground the project in validity and reliability. This is achieved, in part, by a thorough literature review of the state of the field and how it has been researched to date. Chapters 9 and 10 indicated that it is important for a researcher to conduct and report a literature review. A literature review should establish a theoretical framework for the research, indicating the nature and state of the theoretical and empirical fields and important research that has been conducted and policies that have been issued, defining key terms, constructs and concepts, and reporting key methodologies used in other research into the topic. The literature review also sets out what the key issues are in the field to be explored, and why they are, in fact, key issues, and it identifies gaps that need to be plugged in the field. All of this contributes not only to the credibility and validity of the research but to its topicality and significance, and it acts as a springboard into the study, defining the field, what needs to be addressed in it, why, and how it relates to – and extends – existing research in the field. The literature review, then, leads into, and is a foundation for, all areas and stages of the research in question: purpose, foci, research questions, methodology, data analysis, discussion and conclusions.

A literature review may report contentious areas in the field and why they are contentious, contemporary problems that researchers are trying to investigate in the field, difficulties that the field is facing from a research angle, new areas that need to be explored in the field.

A literature review synthesizes several different kinds of materials into an ongoing, cumulative argument that leads to a conclusion (e.g. of what needs to be researched in the present research, how and why). It can be like an extended essay that sets out:

- the argument(s) that the literature review will advance;
- points in favour of the argument(s) or thesis to be advanced/supported;
- points against the argument(s) or thesis to be advanced/supported;
- a conclusion based on the points raised and evidence presented in the literature review.

There are several points to consider in conducting, researching and writing a literature review (cf. University of North Carolina, 2007; Heath, 2009; University of Loughborough, 2009; Creswell, 2012; Wellington, 2015). A literature review:

- defines the field of the research;
- identifies the relevant key concepts, topics, theories, issues, research and ideas in the field under study (including, where relevant, gaps in the field);
- indicates the 'state of the art' in the field chosen;
- sets out the context – temporal, spatial, political etc. – of the research;
- identifies seminal and landmark ideas and research in the field;
- establishes and justifies the need for the research to be conducted, and establishes its significance and originality;
- sets out a rationale for the direction in which the study will go;
- establishes and justifies the methodology to be adopted in the research;
- establishes and justifies the focus of the research;
- sets out and justifies the warrants to be used in the research design.

The literature review is not just a descriptive summary, but an organized and developed argument, usually with subtitles, such that, if the materials were presented in a sequence other than that used, the literature review would lose meaning, coherence, cogency, logic and purpose. It presents, contextualizes, analyses, interprets, critiques and evaluates sources and issues, not just accepting what they say (e.g. it exposes and addresses what the sources overlook, misinterpret, misrepresent, neglect, say something that is contentious, about which they are outdated). It presents arguments and counter-arguments, evidence and counter-evidence about an issue and reveals similarities and differences between authors about the same issue. It sets out and justifies a theoretical framework for the research.

A literature review must state its purposes, methods of working, organization and how it will move to a conclusion, i.e. what it will do, what it will argue, what it will show, what it will conclude and how this links into or informs the subsequent research project. Further, it must state its areas of focus, maybe including a statement of the problem or issue that is being investigated, the hypothesis that the research will test, the themes or topics to be addressed, or the thesis that the research will defend.

A literature review, then, must be conclusive; it must be focused yet comprehensive in its coverage of relevant issues; it must present both sides of an issue or argument; it should address theories, models (where relevant), empirical research, methodological materials, substantive issues, concepts, content and elements of the field in question; and it must include and draw on many sources and types of written material and kinds of data (see, for example, Box 11.2).

In conducting the literature review, Creswell (2012) suggests that the researcher needs to identify key terms, followed by locating the literature, followed by a critical examination of the sources found, for example, for relevance, topicality, accuracy, scope and coverage, followed by the organization of the literature and then subsequent writing of the literature review. For a fuller treatment of conducting and reporting a literature review, we refer readers to Ridley (2010).

A distinction can be drawn between a literature review and a systematic review (cf. Denscombe, 2014). Both collect and synthesize literature, but the former is typically eclectic and even serendipitous, casting its net wide and synthesizing the results, whilst the latter is very focused, typically on empirical research studies (i.e. evidence-based for 'what works'), often those which report research trials (e.g. randomized controlled trials), with stated, often quite narrow or stringent selection and quality criteria, and often requiring measurement and metrics as evidence (though qualitative data are also possible). Systematic reviews are stand-alone documents in their own right, in contrast to literature reviews which tend to be a precursor to an empirical study, clearing the ground for the study to begin. Further, systematic reviews have a narrowly defined scope and focus on a specific question or questions, whereas literature reviews have a wider focus of study.

Systematic reviews typically make explicit the methodologies and criteria they have used in selecting the studies for inclusion (often based on the types and quality of the studies included and their relevance). This is not to argue for literature reviews not being systematic and stringent, or not making clear the criteria used for selecting the literature, or not being rigorous in evaluation of the literature; rather it is to point to the difference in the breadth/narrowness of inclusion criteria and kinds of studies.

Denscombe (2014, pp. 142–3) notes that systematic reviews tend to focus on already-published studies or studies which are publicly available. Whereas in medicine the studies might be of a similar kind (e.g. randomized controlled trials), in the social sciences this may not be the case, rendering comparison and evaluation of studies more problematic (see Chapter 21).

BOX 11.2 TYPES OF INFORMATION IN A LITERATURE REVIEW

Books: hard copy and e-books.

Articles in journals: academic and professional: hard copy and online.

Empirical and non-empirical research.

Reports: from governments, NGOs, organizations, influential associations.

Policy documents: from governments, organizations, 'think tanks'.

Public and private records.

Research papers and reports, for example, from research centres, research organizations.

Theses and dissertations.

Manuscripts.

Databases: searchable collections of records, electronic or otherwise.

Conference papers: local, regional, national, international.

Primary sources: original, first-hand, contemporary source materials such as documents, speeches, diaries and personal journals, letters, autobiographies, memoirs, public records and reports, emails and other correspondence, interview and raw research data, minutes and agendas of meetings, memoranda, proceedings of meetings, communiqués, charters, acts of parliament or government, legal documents, pamphlets, witness statements, oral histories, unpublished works, patents, websites, video or film footage, photographs, pictures and other visual materials, audio-recordings, artefacts, clothing, or other evidence. These are usually produced directly at the time of, close to, or in connection with, the research in question.

Online databases.

Electronic journals or media.

Secondary sources: second-hand, non-original materials, materials written about primary sources, or materials based on sources that were originally elsewhere or which other people have written or gathered, where primary materials have been worked on or with, described, reported, analysed, discussed, interpreted, evaluated, summarized or commented upon, or which are at one remove from the primary sources, or which are written some time after the event, for example, encyclopaedias, dictionaries, newspaper articles, reports, critiques, commentaries, digests, textbooks, research syntheses, meta-analyses, research reviews, histories, summaries, analyses, magazine articles, pamphlets, biographies, monographs, treatises, works of criticism (e.g. literary, political).

Tertiary sources: distillations, collections or compilations of primary and secondary sources, for example, almanacs, bibliographies, catalogues, dictionaries, encyclopaedias, fact books, directories, indexes, abstracts, bibliographies, manuals, guidebooks, handbooks, chronologies.

11.7 Searching for literature on the Internet

The storage and retrieval of research data on the Internet play an important role not only in keeping researchers abreast of developments across the world, but also in providing access to data which can inform literature searches to establish construct and content validity in their own research. Indeed, some kinds of research are essentially large-scale literature searches (e.g. the research papers published in the journals *Review of Educational Research* and *Review of Research in Education*, and materials from the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) at the University of London (http://eppi.ioe.ac.uk/cms) and the What Works Clearinghouse in the United States (http://ies.ed.gov/ncee/wwc)). Online journals, abstracts and titles enable researchers to keep up with the cutting edge of research and to conduct a literature search of relevant material on their chosen topic. Websites and email correspondence enable networks and information to be shared. For example, researchers wishing to gain instantaneous global access to literature and recent developments in research associations can reach all parts of the world in a matter of seconds through websites.

In what follows we indicate the main sources of literature by *kind* only. The companion website to this book gives websites of sources within each kind. Given that websites change and often go out of date quickly, we

strongly recommend that readers go to this companion website, as it is updated and provides many websites, organized by type and source of information. Below we provide websites only for those which have stood the test of time and have not gone out of date for many years.

Researchers wishing to access *educational research associations, organizations and centres* can visit websites such as:

American Educational Research Association: www.aera.net;
Educational Resources Information Center (ERIC): http://eric.ed.gov;
British Educational Research Association: www.bera.ac.uk;
Australian Council for Educational Research: www.acer.edu.au;
European Educational Research Association: www.eera-ecer.de;
National Foundation for Educational Research (UK): www.nfer.ac.uk;
Economic and Social Research Council in the UK: www.esrc.ac.uk.

Researchers wishing to access online journal indices and references for published research results have a variety of websites which they can visit to see *catalogues*, *gateways* and *databases*, and we indicate key sites here on the companion website. These include: the British Education Index; the Organisation for Economic Co-operation and Development (OECD); Social Science Citation Indexes; national statistics services; government departments of education; archives (including statistics databases); the UK's Data Service and Data Archive; UNESCO databases and reports; the Council of European Social Science Data Archive; the gateway to the European Union's sites for data and reports; the United States National Center for Educational Statistics; and the World Bank's gateway to data and statistics.

With regard to searching *libraries*, there are several useful websites for: the British Library and all its online catalogues; the Library site, linking to 18,000 libraries; the United States Library of Congress; the gateway to US libraries; search engines for UK libraries; the Virtual Library; and the Online Computer Library Center. The websites for all these are given in the companion website to this book.

With regard to *items in print*, the website for Books in Print is: www.booksinprint.com, which provides a comprehensive listing of current books in print.

Additional useful *educational research resources* can be found from the National Academies Press (both in total and in its Education Section); centres for the provision of free educational materials and related websites; merged Internet Public Library and the Librarians' Internet Index; and the UK's Research Councils. The websites for all these are given in the companion website to this book.

For *theses*, researchers can go to: the British Library Electronic Theses Online (http://ethos.bl.uk/Home.do); the DART portal for European E-theses; the Aslib Index to Theses; and the Networked Digital Library of Theses and Dissertations (including e-theses). The websites for all of these are given in the companion website to this book.

Most *journals* provide access to abstracts, free online and free alerting services (an email to provide readers with the table of contents of each new issue as it appears), though access to the full article is typically by subscription only. Online journals also provide a comprehensive searching service, in which researchers can search either the specific journal in question or, indeed, the entire range of journals provided by that publisher, using keywords, authors, titles, the digital object identifier (DOI), date and date range, tables of contents, access to articles which appear online before they appear in hard copy etc. Particularly useful here is the facility provided to search the journal in question, or all of that publisher's journals, by keyword. Here the articles can be returned in order by relevance, date, authors, title. It is a first-class facility.

There are many providers of online journals, and we list these, with their websites, in the companion website to the book, including: EBSCO; Emerald Insight; Ingenta; Kluweronline; ProQuest; ProQuest Digital Dissertations and Theses; Science Direct; Web of Knowledge; the Directory of Open Access Journals; the Bath Information and Data Services (BIDS); JSTOR; Journal TOCs (tables of contents). Google Scholar (http://scholar.google.com) is a widely used search engine for articles and books, and it can be interrogated by topic, year, range of years, relevance and the number of citations.

With regard to *statistics*, the companion website to this book provides websites of: the portal to the UK's national statistics; the US National Center for Education Statistics; the UK's Data Service Census Support; and the UK's Office for National Statistics.

When searching the Internet it is useful to keep in mind several points:

- placing words, phrases or sentences inside inverted commas ("…") will keep those words together and in that order in searching for material; this helps to reduce an overload of returned sites;
- placing an asterisk (*) after a word or part of a word will return sites that start with that term but which

have different endings, for example, teach* will return sites on teach, teaching, teacher;
- placing a tilde mark (~) before a word will identify similar words to that which have been entered, for example, ~English teaching will return sites on English language as well as English teaching;
- placing the words *and*, *not*, *or* between phrases or words will return websites where the command indicated in each one of these words is addressed.

Finding research information, where not available from databases and indices on CD-ROMs, is often done through the Internet by trial and error and serendipity, identifying the keywords singly or in combination (between inverted commas). The system of 'bookmarking' websites enables rapid retrieval of these websites for future reference.

Evaluating websites

The use of the Internet for educational research requires an ability to evaluate websites. The Internet is a vast store of disorganized and often unvetted material, and researchers need to be able to ascertain quite quickly how far the web-based material is appropriate. There are several criteria for evaluating websites, including the following (e.g. Tweddle *et al.*, 1998; Rodrigues and Rodrigues, 2000):

- the purpose of the site, as this enables users to establish its relevance and appropriateness;
- the authority and authenticity of the material, which should both be authoritative and declare its sources;
- the content of the material: its up-to-dateness, relevance and coverage;
- the credibility and legitimacy of the material (e.g. is it from a respected source or institution?);
- the correctness, accuracy, completeness and fairness of the material;
- the objectivity and rigour of the material being presented and/or discussed.

In evaluating educational research materials on the web, researchers and teachers can ask themselves several questions:

- Is the author identified?
- Does the author establish her/his expertise in the area, and institutional affiliation?
- Is the organization reputable?
- Is the material referenced; does the author indicate how the material was gathered?
- What is this website designed to do (e.g. to provide information, to persuade)?
- Is the material up-to-date?
- Is the material free from biases, personal opinions and offence?
- How do we know that the author is authoritative on this website?

It is important for the researcher to keep full bibliographic data of the website material used, including the date on which it was retrieved and the website address.

With these preliminary comments, let us turn to the four main areas of the framework for planning research.

11.8 How to operationalize research questions

Chapter 10 indicated that there are many different kinds of research questions that derive from different purposes of the research. For example, research questions may seek:

- to describe what a phenomenon is and what is, or was, happening in a particular situation (e.g. ethnographies, case studies, complexity theory-based studies, surveys);
- to predict what will happen (e.g. experimentation, causation studies, research syntheses);
- to investigate values (e.g. evaluative research, policy research, ideology critique, participatory research);
- to examine the effects of an intervention (e.g. experimentation, ex post facto studies, case studies, action research, causation studies);
- to examine perceptions of what is happening (e.g. ethnography, survey);
- to test a theory;
- to compare the effects of an intervention in different contexts (experimentation, comparative studies);
- to develop, implement, monitor and review an intervention (e.g. participatory research, action research).

Research questions can ask 'what', 'who', 'why', 'when', 'where' and 'how' (cf. Newby, 2010, pp. 65–6). As mentioned in Chapter 10, the researcher has to turn the general purposes of the research into actual practice, i.e. to operationalize the research, turning a general research aim or purpose into specific, particular concrete research questions (or hypotheses) to which exact, specific, concrete answers can be given. It involves specifying a set of operations, elements or behaviours that can be identified, measured or manipulated. The process moves from the general to the particular, from the abstract to the concrete, checking each research question against the research aims until exact, specific, concrete questions have been reached, in all likelihood through an iterative, recursive process (i.e. backwards

and forwards between research aims and emerging research questions) to enable exact, specific, concrete answers to be provided. We provide examples of this in Chapter 10.

11.9 Distinguishing methods from methodologies

In planning research it is important to clarify the distinction between methodology and methods, approaches and instruments, styles of research and ways of collecting data. Simply put, methodology concerns how we find out about the phenomenon, the approach to be used, the principles which underpin it and the justification for using the kind of research approach adopted, the type of study to be conducted, how the research is undertaken (with its associated issues of kinds of research, sampling, instrumentation, canons of validity etc.). Methods concern instrumentation: how data are collected and analysed, whilst methodology justifies the methods used.

The decision on which instrument (method) to use for data collection frequently follows from an earlier decision on which kind (methodology) of research to undertake, for example: a survey; an experiment; an in-depth ethnography; action research; case study research; testing and assessment.

Subsequent chapters of this book set out each of these research styles, their principles, rationales and purposes, and the instrumentation and data types that may be suitable for them. For conceptual clarity it is possible to set out some key features of these (Table 11.3). When decisions have been reached on the stage of research design and methodology, a clear plan of action will have been prepared.

Several of the later chapters of this book are devoted to specific instruments for collecting data, for example: interviews; questionnaires; observation; tests; accounts; biographies; case studies; role-playing; simulations; personal constructs.

11.10 Data analysis

The prepared researcher will need to consider how the data will be analysed. This is important, as it has a specific bearing on the form of the instrumentation. For example, a researcher will need to plan carefully the layout and structure of a questionnaire survey in order to assist data entry for computer reading and analysis; an inappropriate layout may obstruct data entry and subsequent analysis by computer. The planning of data analysis will need to consider:

- What will be done with the data when they have been collected – how will they be processed and analysed?
- How will the results of the analysis be verified, cross-checked and validated?

Decisions will need to be taken with regard to the statistical tests that will be used in data analysis as this will affect the content, type and layout of research items (e.g. in a questionnaire), and the computer packages that are available for processing quantitative and qualitative data, for example, SPSS and NVivo respectively. For statistical processing the researcher will need to ascertain the level of data being processed – nominal, ordinal, interval or ratio (see Chapter 38). Part 5 addresses issues of data analysis and which statistics to use; the choice is not arbitrary (Siegel, 1956; Cohen and Holliday, 1996; Hopkins *et al*., 1996). For qualitative data analysis researchers have at their disposal a range of techniques, for example:

- coding and content analysis of field notes (Miles and Huberman, 1984);
- cognitive mapping (Jones, 1987; Morrison, 1993);
- seeking patterning of responses;
- looking for causal pathways and connections (Miles and Huberman, 1984);
- presenting cross-site analysis (ibid.);
- case studies;
- personal constructs;
- narrative accounts (Flick, 2009; Creswell, 2012);
- action research analysis;
- analytic induction (Denzin, 1989);
- constant comparison and grounded theory (Glaser and Strauss, 1967; Flick 2009; Creswell, 2012);
- discourse analysis (Stillar, 1998);
- biographies and life histories (Atkinson, 1998; Flick, 2009; Creswell, 2012).

The criteria for deciding which forms of data analysis to undertake are governed both by fitness for purpose and legitimacy – the form of data analysis must be appropriate for the kinds of data gathered. For example, it would be inappropriate to use certain statistics with certain kinds of numerical data (e.g. using means with nominal data), or to use causal pathways on unrelated cross-site analysis.

11.11 Presenting and reporting the results

As with the stage of planning data analysis, the prepared researcher will need to consider the form of the

TABLE 11.3 ELEMENTS OF RESEARCH DESIGNS

Model	*Purposes*	*Foci*	*Key terms*	*Characteristics*
Survey	Gathering large-scale data in order to make generalizations Generating statistically manipulable data Gathering context-free data	Opinions Scores Outcomes Conditions Ratings	Measuring Testing Representativeness Generalizability	Describes and explains Represents wide population Gathers numerical data Much use of questionnaires and assessment/test data
Experiment	Comparing under controlled conditions Making generalizations about efficacy Objective measurement of treatment Establishing causality	Initial states, intervention and outcomes Randomized controlled trials	Pre-test and post-test Identification, isolation and control of key variables Generalizations Comparing Causality	Control and experimental groups Treats situations like a laboratory Causes due to experimental intervention Does not judge worth Simplistic
Ethnography	Portrayal of events in subjects' terms Subjective and reporting of multiple perspectives Description, understanding and explanation of a specific situation	Perceptions and views of participants Issues as they emerge over time	Subjectivity Honesty, authenticity Non-generalizable Multiple perspectives Exploration and rich reporting of a specific context Emergent issues	Context-specific Formative and emergent Responsive to emerging features Allows room for judgements and multiple perspectives Wide database gathered over a long period of time Time consuming to process data
Action research	To plan, implement, review and evaluate an intervention designed to improve practice/solve local problem To empower participants through research involvement and ideology critique To develop reflective practice To promote equality democracy To link practice and research To promote collaborative research	Everyday practices Outcomes of interventions Participant empowerment Reflective practice Social democracy and equality Decision making	Action Improvement Reflection Monitoring Evaluation Intervention Problem solving Empowering Planning Reviewing	Context-specific Participants as researchers Reflection on practice Interventionist – leading to solution of 'real' problems and meeting 'real' needs Empowering for participants Collaborative Promoting praxis and equality Stakeholder research

continued

TABLE 11.3 CONTINUED

Case study	To portray, analyse and interpret the uniqueness of real individuals and situations through accessible accounts To catch the complexity and situatedness of behaviour To contribute to action and intervention To present and represent reality – to give a sense of 'being there'	Individuals and local situations Unique instances A single case Bounded phenomena and systems: ■ individual ■ group ■ roles ■ organizations ■ community	Individuality, uniqueness In-depth analysis and portrayal Interpretive and inferential analysis Subjective Descriptive Analytical Understanding specific situations Sincerity Complexity Particularity	In-depth, detailed data from wide data source Participant and non-participant observation Non-interventionist Empathic Holistic treatment of phenomena What can be learned from the particular case
Testing and assessment	To measure achievement and potential To diagnose strengths and weaknesses To assess performance and abilities	Academic and non-academic, cognitive, affective and psychomotor domains – low order to high order Performance, achievement, potential, abilities Personality characteristics	Reliability Validity Criterion-referencing Norm-referencing Domain-referencing Item-response Formative Summative Diagnostic Standardization Moderation	Materials designed to provide scores that can be aggregated Enables individuals and groups to be compared In-depth diagnosis Measures performance

reporting of the research and its results, giving due attention to the needs of different audiences (e.g. an academic audience may require different contents from a wider professional audience and, a fortiori, from a lay audience). Decisions here address:

- How to write up and report the research;
- When to write up and report the research (e.g. ongoing or summative);
- How to present the results in tabular and/or written-out form;
- How to present the results in non-verbal forms;
- To whom to report (the necessary and possible audiences of the research);
- How frequently to report.

For an example of setting out a research report, see the accompanying website.

11.12 A planning matrix for research

In planning a piece of research, the range of questions to be addressed can be set into a matrix. Table 11.4 provides such a matrix, in the left-hand column of which are the questions which figure in the four main areas set out so far:

1 orienting decisions;
2 research design and methodology;
3 data analysis;
4 presenting and reporting the results.

Questions 1–10 are the orienting decisions, questions 11–22 concern the research design and methodology, questions 23–4 cover data analysis, and questions 25–30 deal with presenting and reporting the results. Within each of the thirty questions there are several sub-questions which research planners may need to address. For example, within question 5 ('What are the

TABLE 11.4 A MATRIX FOR PLANNING RESEARCH

Orienting decisions

Question	*Sub-issues and problems*	*Decisions*
1 Who wants the research?	Is the research going to be useful? Who might wish to use the research? Are the data going to be public? What if different people want different things from the research? Can people refuse to participate?	Find out the controls over the research which can be exercised by respondents. Set out the scope and audiences of the research. Determine the reporting mechanisms.
2. Who will receive the research?	Will participants be able to veto the release of parts of the research to specified audiences? Will participants be able to give the research to whomsoever they wish? Will participants be told to whom the research will go?	Determine the proposed internal and external audiences of the research. Determine the controls over the research which can be exercised by the participants. Determine the rights of the participants and the researcher to control the release of the research.
3. What powers do the recipients of the research have?	What use will be made of the research? How might the research be used for or against the participants? What might happen if the data fall into the 'wrong' hands? Will participants know in advance what use will and will not be made of the research?	Determine the rights of recipients to do what they wish with the research. Determine the respondents' rights to protection as a result of the research.
4 What are the timescales of the research?	Is there enough time to do all the research? How to decide what to be done within the timescale?	Determine the timescales and timing of the research.
5 What are the purposes of the research?	What are the formal and hidden agendas here? Whose purposes are being served by the research? Who decides the purposes of the research? How will different purposes be served in the research?	Determine all the possible uses of the research. Determine the powers of the respondents to control the uses made of the research. Decide on the form of reporting and the intended and possible audiences of the research.
6 What are the research questions?	Who decides what the questions will be? Do participants have rights to refuse to answer or take part? Can participants add their own questions?	Determine the participants' rights and powers to participate in the planning, form and conduct of the research. Decide the balance of all interests in the research.
7 What must be the focus in order to answer the research questions?	Is sufficient time available to focus on all the necessary aspects of the research? How will the priority foci be decided? Who decides the foci?	Determine all the aspects of the research, prioritize them, and agree on the minimum necessary areas of the research. Determine decision-making powers on the research.

continued

TABLE 11.4 CONTINUED

8 What costs are there – human, material, physical, administrative, temporal?	What support is available for the researcher? What materials are necessary?	Cost out the research.
9 Who owns the research?	Who controls the release of the report? What protections can be given to participants? Will participants be identified and identifiable/traceable? Who has the ultimate decision on what data are included?	Determine who controls the release of the report. Decide the rights and powers of the researcher. Decide the rights of veto. Decide how to protect those who may be identified/identifiable in the research.
10 At what point does the ownership pass from the respondent to the researcher and from the researcher to the recipients?	Who decides the ownership of the research? Can participants refuse to answer certain parts if they wish, or, if they have the option not to take part, must they opt out of everything? Can the researcher edit out certain responses?	Determine the ownership of the research at all stages of its progress. Decide the options available to the participants. Decide the rights of different parties in the research, e.g. respondents, researcher, recipients.

Research design and methodology

Question	*Sub-issues and problems*	*Decisions*
11 What are the specific purposes of the research?	How do these purposes derive from the overall aims of the research? Will some areas of the broad aims be covered, or will the specific research purposes have to be selective? What priorities are there?	Decide the specific research purposes and write them as concrete questions.
12 How are the general research purposes and aims operationalized into specific research questions?	Do the specific research questions together cover all the research purposes? Are the research questions sufficiently concrete as to suggest the kinds of answers and data required and the appropriate instrumentation and sampling? How to balance adequate coverage of research purposes with the risk of producing an unwieldy list of sub-questions?	Ensure that each main research purpose is translated into specific, concrete questions that, together, address the scope of the original research questions. Ensure that the questions are sufficiently specific as to suggest the most appropriate data types, kinds of answers required, sampling and instrumentation. Decide how to ensure that any selectivity still represents the main fields of the research questions.
13 What are the specific research questions?	Do the specific research questions demonstrate construct and content validity?	Ensure that the coverage and operationalization of the specific questions addresses content and construct validity respectively.
14 What needs to be the focus of the research in order to answer the research questions?	How may foci are necessary? Are the foci clearly identifiable and operationalizable?	Decide the number of foci of the research questions. Ensure that the foci are clear and can be operationalized.

15 What is the main methodology of the research?	How many methodologies are necessary? Are several methodologies compatible with each other? Will a single focus/research question require more than one methodology (e.g. for triangulation and concurrent validity)?	Decide the number, type and purposes of the methodologies to be used. Decide whether one or more methodologies is/are necessary to gain answers to specific research questions. Ensure that the most appropriate form of methodology is employed.
16 How will validity and reliability be addressed?	Will there be the opportunity for cross-checking? Will the depth and breadth required for content validity be feasible within the constraints of the research (e.g. time constraints, instrumentation)? In what senses are the research questions valid (e.g. construct validity)? Are the questions fair? How does the researcher know if people are telling the truth? What kinds of validity and reliability are to be addressed? How will the researcher take back the research to respondents for them to check that the interpretations are fair and acceptable? How will data be gathered consistently over time? How to ensure that each respondent is given the same opportunity to respond?	Determine the process of respondent validation of the data. Decide a necessary minimum of topics to be covered. Subject the plans to scrutiny by critical friends ('jury' validity). Pilot the research. Build in cross-checks on data. Address the appropriate forms of reliability and validity. Decide the questions to be asked and the methods used to ask them. Determine the balance of open and closed questions.
17 How will reflexivity be addressed?	How will reflexivity be recognized? Is reflexivity a problem? How can reflexivity be included in the research?	Determine the need to address reflexivity and to make this public. Determine how to address reflexivity in the research.
18 What kinds of data are required?	Does the research need words, numbers or both? Does the research need opinions, facts or both? Does the research seek to compare responses and results or simply to illuminate an issue?	Determine the most appropriate types of data for the foci and research questions. Balance objective and subjective data. Determine the purposes of collecting different types of data and the ways in which they can be processed.
19 From whom will data be acquired (i.e. sampling)?	Will there be adequate time to go to all the relevant parties? What kind of sample is required (e.g. probability/non-probability/random/ stratified etc.)? How to achieve a representative sample (if required)?	Determine the minimum and maximum sample. Decide on the criteria for sampling. Decide the kind of sample required. Decide the degree of representativeness of the sample. Decide how to follow up and not to follow up on the data gathered.

continued

TABLE 11.4 CONTINUED

20 Where else will data be available?	What documents and other written sources of data can be used? How to access and use confidential material? What will be the positive or negative effects on individuals of using certain documents?	Determine the necessary/desirable/possible documentary sources. Decide access and publication rights and protection of sensitive data.
21 How will the data be gathered (i.e. instrumentation)?	What methods of data gathering are available and appropriate to yield data to answer the research questions? What methods of data gathering will be used? How to construct interview schedules/questionnaires/tests/ observation schedules? What will be the effects of observing participants? How many methods should be used (e.g. to ensure reliability and validity)? Is it necessary or desirable to use more than one method of data collection on the same issue? Will many methods yield more reliable data? Will some methods be unsuitable for some people or for some issues?	Determine the most appropriate data-collection instruments to gather data to answer the research questions. Pilot the instruments and refine them subsequently. Decide the strengths and weaknesses of different data-collection instruments in the short and long term. Decide which methods are most suitable for which issues. Decide which issues will require more than one data-collection instrument. Decide whether the same data-collection methods will be used with all the participants.
22 Who will undertake the research?	Can different people plan and carry out different parts of the research?	Decide who will carry out the data collection, processing and reporting.
Data Analysis		
Question	*Sub-issues and problems*	*Decisions*
23 How will the data be analysed?	Are the data to be processed numerically or verbally? What computer packages are available to assist data processing and analysis? What statistical tests will be needed? How to perform a content analysis of word data? How to summarize and present word data? How to process all the different responses to open-ended questions? Will the data be presented person by person, issue by issue, aggregated to groups, or a combination of these? Does the research seek to make generalizations? Who will process the data?	Clarify the legitimate and illegitimate methods of data processing and analysis of quantitative and qualitative data. Decide which methods of data processing and analysis are most appropriate for which types of data and for which research questions. Check that the data processing and analysis will serve the research purposes. Determine the data protection issues if data are to be processed by 'outsiders' or particular 'insiders'.

24 How to verify and validate the data and their interpretation?	What opportunities will there be for respondents to check the researcher's interpretation? At what stages of the research is validation necessary? What will happen if respondents disagree with the researcher's interpretation?	Determine the process of respondent validation during the research. Decide the reporting of multiple perspectives and interpretations. Decide respondents' rights to have their views expressed or to veto reporting.
Presenting and reporting the results		
Question	*Sub-issues and problems*	*Decisions*
25 How to write up and report the research?	Who will write the report and for whom? How detailed must the report be? What must the report contain? What channels of dissemination of the research are to be used?	Ensure that the most appropriate form of reporting is used for the audiences. Keep the report as short, clear and complete as possible. Provide summaries if possible/fair. Ensure that the report enables fair critique and evaluation to be undertaken.
26 When to write up and report the research (e.g. ongoing or summative)?	How many times are appropriate for reporting? For whom are interim reports compiled? Which reports are public?	Decide the most appropriate timing, purposes and audiences of the reporting. Decide the status of the reporting (e.g. formal, informal, public, private).
27 How to present the results in tabular and/or written-out form?	How to ensure that everyone will understand the language or the statistics? How to respect the confidentiality of the participants? How to report multiple perspectives?	Decide the most appropriate form of reporting. Decide whether to provide a glossary of terms. Decide the format(s) of the reports. Decide the number and timing of the reports. Decide the protection of the individual's rights, balancing this with the public's rights to know.
28 How to present the results in non-verbal forms?	Will different parties require different reports? How to respect the confidentiality of the participants? How to report multiple perspectives?	Decide the most appropriate form of reporting. Decide the number and timing of the reports. Ensure that a written record is kept of oral reports. Decide the protection of the individual's rights, balancing this with the public's rights to know.
29 To whom to report (the necessary and possible audiences of the research)?	Do all participants receive a report? What will be the effects of not reporting to stakeholders?	Identify the stakeholders. Determine the least and most material to be made available to the stakeholders.
30 How frequently to report?	Is it necessary to provide interim reports? If interim reports are provided, how might this affect the future reports or the course of the research?	Decide on the timing and frequency of the reporting. Determine the formative and summative nature of the reports.

purposes of the research?') the researcher would have to differentiate major and minor purposes, explicit and maybe implicit purposes, whose purposes are being served by the research and whose interests are being served by the research. An example of these sub-issues and problems is contained in the second column.

At this point the planner is still at the divergent phase of the research planning, dealing with *planned possibilities*, opening up the research to all facets and interpretations. In the column headed 'decisions' the research planner is moving towards a convergent phase, where planned possibilities become visible within the terms of constraints available to the researcher. Here the researcher moves down the column marked 'decisions' to see how well the decision which is taken in regard to one issue/question fits in with the decisions in regard to other issues/questions. For one decision to fit with another, four factors must be present:

1. All of the cells in the 'decisions' column must be coherent – they must not contradict each other;
2. All of the cells in the 'decisions' column must be mutually supporting;
3. All of the cells in the 'decisions' column must be practicable when taken separately;
4. All of the cells in the 'decisions' column must be practicable when taken together.

Not all of the planned possibilities might be practicable when these four criteria are applied. It would be of very little use if the methods of data collection listed in the 'decisions' column of question 21 ('How will the data be gathered?') offered little opportunity to fulfil the needs of acquiring information to answer question 7 ('What must be the focus in order to answer the research questions?'), or if the methods of data collection are impracticable within the timescales available in question 4.

In the matrix of Table 11.4 the cells have been completed in a deliberately content-free way, i.e. the matrix as presented here does not deal with the specific, actual points which might emerge in a particular research proposal. If the matrix were to be used for planning an actual piece of research, then, instead of couching the wording of each cell in generalized terms, it would be more useful if *specific, concrete* responses were given which address particular issues and concerns in the research proposal.

Many of these questions concern rights, responsibilities and the political uses (and abuses) of the research. This underlines the view that research is an inherently political and moral activity; it is not politically or morally neutral. The researcher has to be concerned with the uses as well as the conduct of the research.

11.13 Managing the planning of research

It should not be assumed that research will always go according to plan. For example, the attrition of the sample might happen (participants leaving during the research), or a poor response rate to questionnaires might be encountered, rendering subsequent analysis, reporting and generalization problematical; administrative support might not be forthcoming, or there might be serious slippage in the timing. This is not to say that a plan for the research should not be made; rather it is to suggest that it is dangerous to put absolute faith in it. For an example of what to include in a research proposal, see the accompanying website.

To manage the complexity in planning outlined above, a simple four-stage model can be proposed:

Stage 1: Identify the purposes of the research.
Stage 2: Identify and give priority to the constraints under which the research will take place;
Stage 3: Plan the possibilities for the research within these constraints.
Stage 4: Decide the research design.

Each stage contains several operations. Figure 11.1 clarifies this four-stage model, drawing out the various operations contained in each stage.

Research planners can consider which instruments will be used at which stage of the research and with which sectors of the sample population. Table 11.5 sets out a matrix of these for planning, for example, a small-scale piece of research.

A matrix approach such as this enables research planners to see at a glance their coverage of the sample and of the instruments used at particular points in time, making omissions clear and promoting such questions as:

- Why are certain instruments used at certain times and not at others?
- Why are certain instruments used with certain people and not with others?
- Why do certain times in the research use more instruments than other times?
- Why is there such a concentration of instruments at the end of the study?
- Why are certain groups involved in more instruments than other groups?
- Why are some groups apparently neglected (e.g. parents), for example, is there a political dimension to the research?
- Why are questionnaires the main kinds of instrument to be used?

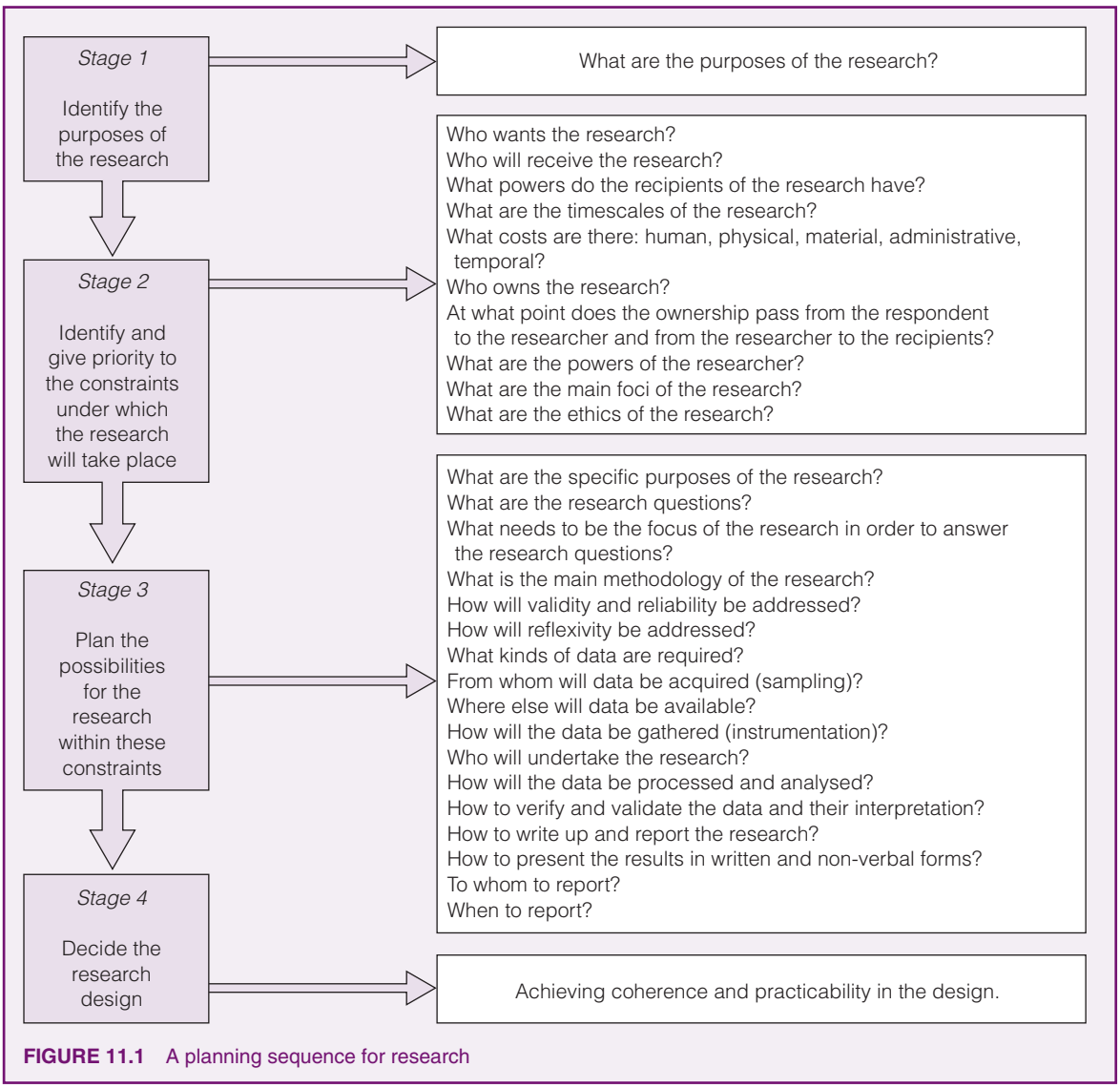


FIGURE 11.1 A planning sequence for research

- Why are some instruments (e.g. observation, testing) not used at all?
- What makes the five stages separate?
- Are documents only held by certain parties (and, if so, might one suspect an 'institutional line' to be revealed in them)?
- Are some parties more difficult to contact than others (e.g. university teacher educators)?
- Are some parties more important to the research than others (e.g. school principals)?
- Why are some parties excluded from the sample (e.g. school governors, policy makers, teachers' associations and unions)?
- What is the difference between the three groups of teachers?

Matrix planning is useful for exposing key features of the planning of research. Further matrices might be constructed to indicate other features of the research, for example:

- the timing of the identification of the sample;
- the timing of the release of interim reports;
- the timing of the release of the final report;
- the timing of pre-tests and post-tests (in an experimental style of research);

TABLE 11.5 A PLANNING MATRIX FOR RESEARCH

Time sample	*Stage 1 (start)*	*Stage 2 (3 months)*	*Stage 3 (6 months)*	*Stage 4 (9 months)*	*Stage 5 (12 months)*
Principal/ Headteacher	Documents Interview Questionnaire 1	Interview	Documents Questionnaire 2	Interview	Documents Interview Questionnaire 3
Teacher group 1	Questionnaire 1		Questionnaire 2		Questionnaire 3
Teacher group 2	Questionnaire 1		Questionnaire 2		Questionnaire 3
Teacher group 3	Questionnaire 1		Questionnaire 2		Questionnaire 3
Students			Questionnaire 2		Interview
Parents	Questionnaire 1		Questionnaire 2		Questionnaire 3
University teacher educators	Interview Documents				Interview Documents

- the timing of intensive necessary resource support (e.g. reprographics);
- the timing of meetings of interested parties.

These examples cover timings only; other matrices might be developed to cover other combinations, for example: reporting by audiences; research team meetings by reporting; instrumentation by participants etc. They are useful summary devices.

11.14 A worked example

Let us say that a school is experiencing low morale and the researcher has been brought in to investigate the school's organizational culture as it impacts on morale. The researcher has been given open access to the school and has five months from the start of the project to producing the report. (For a fuller version of this, see the accompanying website.) She plans the research thus:

1 Purposes

i To present an overall and in-depth picture of the organizational culture(s) and subcultures, including the prevailing cultures and subcultures, within the school;
ii To provide an indication of the strength of the organizational culture(s);
iii To make suggestions and recommendations about the organizational culture of, and its development at, the school.

2 Research questions

i What are the major and minor elements of organizational culture in the school?
ii What are the organizational cultures and subcultures in the school?
iii Which (sub)cultures are the most and least prevalent in the school, and in which parts of the school are these most and least prevalent?
iv How strong and intense are the (sub)cultures in the school?
v What are the causes and effects of the (sub)cultures in the school?
vi How can the (sub)cultures be improved in the school?

3 Focus

Three levels of organizational cultures will be examined:

i underlying values and assumptions;
ii espoused values and enacted behaviours;
iii artefacts.

Organizational culture concerns values, assumptions, beliefs, espoused theories, observed practices, areas of conflict and consensus, the formal and hidden messages contained in artefacts, messages, documents and language, the 'way we do things', the physical environment, relationships, power, control, communication, customs and rituals, stories, the reward system and motivation, the micro-politics of the school, involvement in decision making, empowerment and exploitation/manipulation, leadership, commitment, and so on.

In terms of the 'possible sequence of considerations' set out earlier in the chapter, the 'preparatory issues' here include: (i) a literature review on organizational culture, organizational health, leadership of organizations, motivation, communication and empowerment; (ii) the theoretical framework underpinning the research

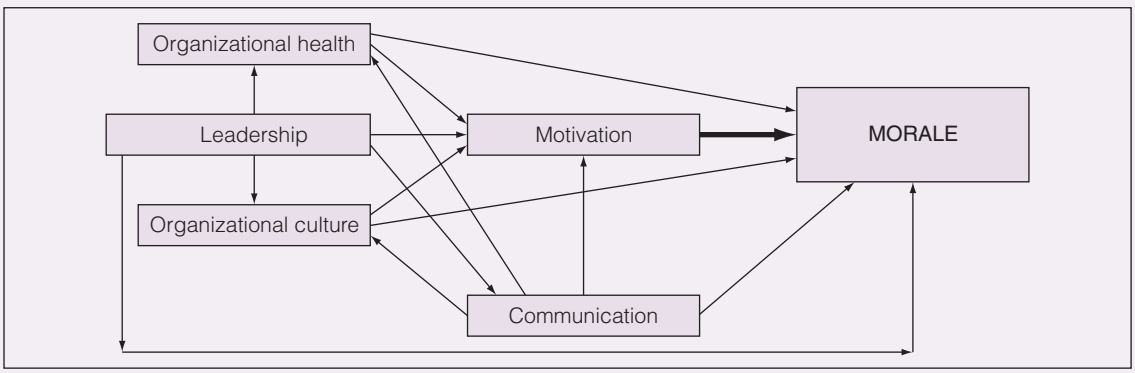
(see Figure 11.2); and (iii) the devising of the conceptual framework to include: levels of organizational culture (artefacts, enacted values and underlying assumptions; see Figure 11.3); key features of organizational health; key issues in, and styles of, leadership; key features of communication (e.g. direction, content, medium); and motivation (intrinsic and extrinsic). Together these constitute the ontological dimension of the 'preparatory issues' of the 'possible sequence of considerations'.

4 Methodology

The methodologies here address the epistemological dimension of the 'preparatory issues' of the 'possible sequence of considerations' set out earlier in the chapter: how we can know about, and research, the phenomenon. Here organizational culture is intangible, yet its impact on a school's operations and morale is very tangible. This suggests that, whilst quantitative measures may be used, they are likely only to yield comparatively superficial information about the school's culture. In order to probe beneath the surface of the school's culture, to examine the less overt aspects of the school's culture(s) and subcultures, it is important to combine quantitative and qualitative methodologies for data collection. A mixed methodology will be used for the data collection, using numerical and verbal data, in order to gather rounded, reliable data. A survey approach will be used to gain an overall picture, and a more fine-grained analysis will be achieved through qualitative approaches (Figure 11.3).

5 Instrumentation

The data gathered will be largely perception-based, and will involve gathering employees' views of the (sub)cultures. As the concept of organizational culture is derived,



Though at first sight the graphic looks complex, because there are many arrows, in fact it is not complicated. The theory underpinning this, which derives from a literature review of empirical studies of organizational behaviour, leadership, individual and social psychology, is that these five identified key factors influence morale: organizational health, organizational culture, leadership, communication and motivation. Of course, there are many, many more factors, but the research has assumed that these are key factors in the present study. This highlights an important feature of theory: it is selective in what it includes and it operates at a high level of generality (a conceptual model would provide much closer detail here, breaking down the main areas into more specific elements).

The arrows indicate the assumed directions of influence of key factors in morale which derive from literature. Here *organizational health* and *organizational culture* have a direct effect on morale and motivation; *leadership* has a direct effect on organizational health, organizational culture, motivation, communication and morale – in other words it is a key factor; *communication* has a direct effect on motivation, organizational culture, organizational health and morale – in other words, it is an important factor; and *motivation* has a direct effect on morale. Note that the direction of inferred causality is one-way, even though, in reality, the causality is multi-directional and reciprocal. This indicates another key feature of the theory: it is selective in its inferred or assumed direction of causality (and, indeed, in causal modelling).

The theory here is also that leadership is a key driver: note that the causal arrows lead *from*, rather than *to*, leadership. Further, motivation is a key recipient of factors, and, in turn, it is assumed to influence morale. One can infer from this that motivation exerts an important influence on morale, and this is reflected in the thickness of the causal arrow from motivation to morale.

The graphic here, then, is a portrayal of the theoretical assumptions that underpin the research on morale.

FIGURE 11.2 Theoretical framework for investigating low morale in an organization

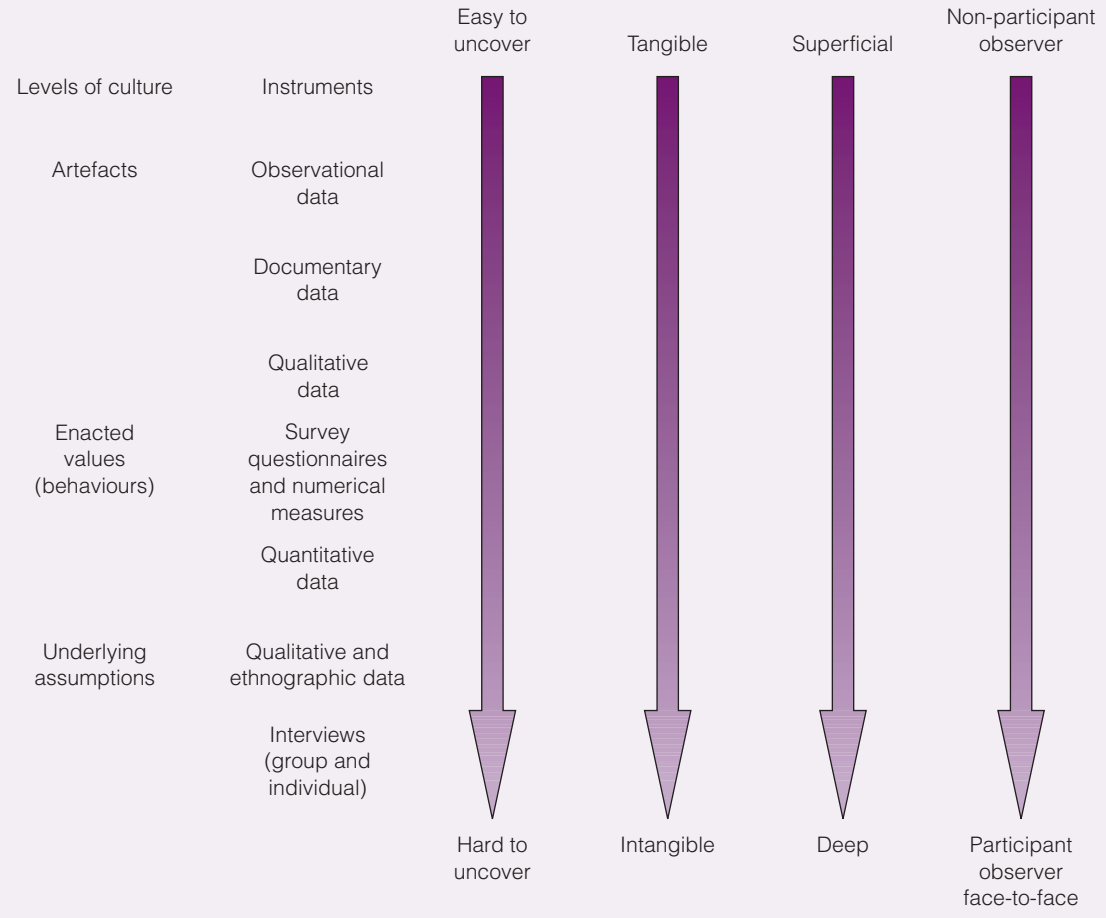


FIGURE 11.3 Understanding the levels of organizational culture

in part, from ethnography and anthropology, the research will use qualitative and ethnographic methods.

One of the difficulties anticipated is that the less tangible aspects of the school might be the most difficult on which to collect data. Not only will people find it harder to articulate responses and constructs, but they may also be reluctant to reveal these in public. The more the project addresses intangible and unmeasurable elements, and the richer the data that are to be collected, the more there is a need for increased and sensitive interpersonal behaviour, face-to-face data-collection methods, and qualitative data.

There are several instruments for data collection: questionnaires, semi-structured interviews (individual and group), observational data, documentary data and reports will constitute a necessary minimum, as follows (see also Figure 11.3):

i *Questionnaire surveys*, using commercially available instruments, each of which measures different aspects of school's culture, in particular:

- the organizational culture questionnaire by Harrison and Stokes (1992), which looks at overall cultures and provides a general picture in terms of *role*, *power*, *achievement* and *support* cultures, and examines the differences between existing and preferred cultures;
- the Organizational Culture Inventory by Cooke and Lafferty (1989), which provides a comprehensive and reliable analysis of the presenting organizational cultures.

Questionnaires, using rating scales, will catch articulated, espoused, enacted, visible aspects of organizational

culture, and will measure, for example, the extent of sharedness of culture, congruence between existing and ideal, and strength and intensity of culture.

ii *Semi-structured qualitative interviews for individuals and groups*, gathering data on the more intangible aspects of the school's culture, for example, values, assumptions, beliefs, wishes, problems. Interviews will be semi-structured, i.e. with a given agenda and open-ended questions. As face-to-face individual interviews might be intimidating for some groups, group interviews will be used. In all of the interviews the important part will be the supplementary question, 'why?'.

iii *Observational data* will comment on the physical environment, and will then be followed up with interview material to discover participants' responses to, perceptions of, messages contained in, and attitudes to, the physical environment. Artefacts, clothing, shared and private spaces, furniture, notices, regulations etc. all give messages to participants.

iv *Documentary analysis and additional stored data*, reporting the formal matters in the school, examined for what they include and what they exclude.

6 Sampling

i The questionnaire will be given to all employees who are willing to participate;

ii The semi-structured interviews will be conducted on a 'critical case' basis, i.e. with participants who are in key positions and who are 'knowledgeable people' about the activities and operations of the school.

There will be stratified sampling for the survey instruments, in order to examine how perceptions of the school's organizational culture vary according to the characteristics of the sub-samples. This will enable the levels of congruence or disjunction between the responses of the various sub-groups to be charted. Nominal characteristics of the sampling will be included, for example, age, level in the school, departments, gender, ethnicity, nationality and years of working in the school.

7 Parameters

i The data will be collected on a 'one-shot' basis rather than longitudinally;

ii A multi-method approach will be used for data collection.

8 Stages in the research

There are five stages in the research:

Stage one: Development and operationalization, including

i A review of literature and commercially produced instruments;

ii Clarification of the research questions;

iii Clarification of methodology and sampling;

Stage two: Instrumentation and the piloting of the instruments

i Questionnaire development and piloting;

ii Semi-structured interview schedules and piloting;

iii Gathering of observational data;

iv Analysis of documentary data;

Because of the limited number of senior staff, it will not be possible to conduct pilot interviews with them, as this will preclude them from the final data collection.

Stage three: Data collection, which will proceed in the following sequence

Administration of the questionnaire → Analysis of questionnaire data to provide material for the interviews → Interviews to be conducted concurrently.

Stage four: Data analysis and interpretation

Numerical data will be processed with SPSS, which will also enable the responses from sub-groups of the school to be separated for analysis. Qualitative data will be analysed using protocols of content analysis.

Stage five: Reporting

A full report on the findings will include conclusions, implications and recommendations.

9 Ethics and ownership

Participation in the project will be on the basis of informed consent, and on a voluntary basis, with rights of withdrawal at any time. Given the size and scope of the cultural survey, it is likely that key people in the school will be identifiable, even though the report is confidential. This will be made clear to the potential participants. Copies of the report will be available for all the employees. Data, once given to the researcher, are his/hers, and she/he may not use them in any way which will publicly identify the school; the report is the property of the school.

10 Time frames

The project will be completed in five months:

BOX 11.3 A CHECKLIST FOR PLANNING RESEARCH

1. How have you taken account of the ontological and epistemological characteristics of the phenomenon to be investigated?
2. Have you clarified the purposes of the research?
3. What do you want the research to do, to 'deliver', to find out?
4. What are the purposes and objectives of the research?
5. Have you identified the constraints on your research? What are they?
6. Is your research feasible within the required time frames?
7. What approaches to the research (methodologies) are most suitable for the research, in terms of the ontology and epistemology of the phenomenon under investigation, and the purposes of the research?
8. What warrants have you provided to link evidence to conclusions?
9. What are the methodology(ies) and paradigm(s) on which the research is built? How comfortably do they fit the research purposes and the nature of the phenomena under investigation?
10. Does your research seek to test a theory or hypothesis, to develop a theory, to investigate and explore, to understand, to describe, to develop specific practices, to evaluate, to investigate?
11. Will your research best be accomplished by research that is naturalistic, interpretive, positivist, post-positivist, mixed methods-based, participatory, evaluatory, ideology critical, feminist, complexity theory-based, either alone or in combination?
12. Will your research use survey, documentary research, quantitative methods, ethnographic or qualitative methods, experiments, historical sources, action research, case studies, *ex post facto* designs, either alone or in combination?
13. Do you need to identify independent and dependent variables?
14. Is your research seeking to establish causation?
15. Are you seeking to generalize from your research?
16. In planning your research, have you indicated how you will address validity and reliability in the conceptualization, planning, methodology, instrumentation, data analysis, discussion, the drawing of conclusions and reporting?
17. Who will gather, enter, process, analyse, interpret and verify your data?
18. Have you identified how you will address reflexivity?
19. Have you identified what you need to focus on in order to answer the research questions and conduct the research?
20. Have you identified whom you need to contact in connection with conducting the research?
21. Have you checked that all the ethical issues in the research have been addressed with all the necessary parties? Have you gained ethical clearance to conduct the research?
22. Is your research overt or covert? If it is covert, or involves intentional deceit, how is this justified?
23. Have you conducted a literature review, and how does the literature review inform your research?
24. Does your research need research questions? If not, why not? If so, what are they and have they been operationalized comprehensively, concretely and fairly?
25. Have you operationalized your research purposes into research questions?
26. What are the timescales for the different stages of your research?
27. Have you identified what kinds of data you need at different stages of the research, and why?
28. Have you identified the instruments that you will need for data collection at the different stages of the research, for example: interviews, questionnaires, observations, role-plays, accounts, personal constructs, tests, case studies, field notes, diaries, documents, etc.?
29. Is your research 'front-loaded' or 'end-loaded' in terms of planning, conduct and analysis?
30. Who are the participants?
31. Do you need a sample or a population? What is the population and what are the sample and the sampling strategy?
32. Have you planned how you will analyse the data, and at what stages of the research?
33. Have you planned how you will validate your data and interpretation of the data?
34. Have you planned when and how you will report and present the research findings, and to whom?
35. Have you planned how you will disseminate your research findings?
36. Have you identified what controls you will place on the release of your findings, and to whom, why and for how long, and who owns the research and the data?

- the first month for a review of the relevant literature;
- the second month to develop the instrumentation and research design;
- the third month to gather the data;
- the fourth month to analyse the data;
- the fifth month to complete the report.

The example indicates a systematic approach to the planning and conduct of the research that springs from a perceived need in the school. It works within given constraints and makes clear what it will 'deliver'. Though the research does not specify hypotheses to be tested, nevertheless it would not be difficult to convert the research questions into hypotheses if this style of research were preferred.

11.15 Ensuring quality in the planning of research

'Fitness for purpose' reigns in planning research; the research plan must suit the purposes of the research. If the reader is left feeling, at the end of this chapter, that the task of research is complex, then that is an important message, for rigour and thoughtful, thorough planning are necessary if the research is to be worthwhile and effective. For a checklist for evaluating research, see Box 11.3 and the accompanying website.

The intention of the research planning and design is to ensure that rigour, fitness for purpose and high quality are addressed. Furlong and Oancea (2005, pp. 11–15) identify several clear dimensions of quality in educational research. For theoretical and methodological robustness they identify quality in terms of: (a) the 'trustworthiness' of the research; (b) its 'contribution to knowledge'; (c) its 'explicitness in designing and reporting'; (d) its 'propriety' (conformance to legal and ethical requirements); and (e) the 'paradigm-dependence' (fidelity to the paradigm, ontology and epistemological premises of the research).

For 'value for use' (the 'technological dimension'), Furlong and Oancea (2005, pp. 12–13) identify key indicators of quality as: (a) the 'salience/timeliness' of the research; (b) its 'purposivity' (fitness for purpose); (c) its 'specificity and accessibility' (scope, responsiveness to user needs, and predicted usage); (d) its 'concern for enabling impact' (dissemination for impact); and (e) its 'flexibility and operationalisability' (development into practical terms and utility for audiences).

For 'capacity building and value for people' (Furlong and Oancea, 2005, pp. 13–14), they identify key indicators of quality as residing in: (a) 'partnership, collaboration and engagement'; (b) 'plausibility' ('from the practitioner's perspective'); (c) 'reflection and criticism' (research that develops reflexivity and self-reflection); (d) 'receptiveness' (research that enhances the receptiveness of practitioners and a wider audience); and (e) 'stimulating personal growth'.

For their 'economic dimension', Furlong and Oancea (2005, pp. 14–15) indicate six elements of quality in research: (a) 'cost-effectiveness'; (b) 'marketability' and 'competitiveness' (e.g. in the research market); (c) 'auditability'; (d) 'feasibility'; (e) 'originality'; and (f) 'value-efficiency'.

The sections of this chapter and the preceding chapter, separately and together, have indicated how these can be addressed in the planning of research.

Companion Website

The companion website to the book provides PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Sampling

CHAPTER 12

Sampling is a crucial element of research, and this chapter introduces key issues in sampling, including:

- sample size
- statistical power
- sampling error
- sample representativeness
- access to the sample
- sampling strategy
- probability samples
- non-probability samples
- sampling in qualitative research
- sampling in mixed methods research
- planning the sampling

12.1 Introduction

The quality of a piece of research stands or falls by the appropriateness of its methodology and instrumentation and by the suitability of the sampling strategy that has been adopted. Questions of sampling arise directly out of the issue of defining the population on which the research will focus.

Researchers must take sampling decisions early in the overall planning of research, not least of which is whether to have a sample or an entire population. However, as this chapter concerns sampling we keep to this topic, and here factors such as expense, time and accessibility frequently prevent researchers from gaining information from the whole population. Therefore they often need to be able to obtain data from a smaller group or subset of the total population in such a way that the knowledge gained is representative of the total population (however defined) under study. This smaller group or subset is the *sample*. Experienced researchers start with the total population and work down to the sample. By contrast, less experienced researchers often work from the bottom up, that is, they determine the minimum number of respondents needed to conduct the research (Bailey, 1994). However, unless they identify the total population in advance, it is virtually impossible for them to assess how representative the sample is that they have drawn.

Suppose that a class teacher has been released from her teaching commitments for one month in order to conduct some research into the abilities of thirteen-year-old students to undertake a set of science experiments. The research is to draw on three secondary schools which contain 300 such students each, a total of 900 students, and the method that the teacher has been asked to use for data collection is a semi-structured interview. Because of the time available to the teacher it would be impossible for her to interview all 900 students (the total population being all the cases). Therefore she has to be selective and to interview fewer than all 900 students. How will she decide that selection; how will she select which students to interview?

If she were to interview 200 of the students, would that be too many? If she were to interview just twenty of the students, would that be too few? If she were to interview just the males or just the females, would that give her a fair picture? If she were to interview just those students whom the science teachers had decided were 'good at science', would that yield a true picture of the total population of 900 students? Perhaps it would be better for her to interview those students who were experiencing difficulty in science and who did not enjoy science, as well as those who were 'good at science'. Suppose that she turns up on the days of the interviews only to find that those students who do not enjoy science have decided to absent themselves from the science lesson. How can she reach those students?

Decisions and problems such as these face researchers in deciding the sampling strategy to be used. Judgements have to be made about several key factors in sampling, for example:

- the sample size;
- statistical power;
- the representativeness and parameters of the sample;
- access to the sample;
- the sampling strategy to be used;
- the kind of research that is being undertaken (e.g. quantitative/qualitative/mixed methods).

The decisions here will influence the sampling strategy to be used. This assumes that a sample is actually required; there may be occasions on which the researcher can access the whole population rather than a sample.

Uprichard (2013) adds to these a range of ontological, epistemological and logistical matters. Ontological matters concern the unit of analysis – why choose the unit of analysis (the 'case') that has been chosen? For example, a key problem in addressing populations and samples is whether the population size and characteristics are actually known (which are needed to identify a sampling frame) and how much we know about the sample, and this may be a major difficulty in some kinds of social and educational research (Uprichard, 2013, p. 3). This is also an ontological and epistemological problem, for example, how we have any knowledge of the population and the sample (the 'cases') and what that knowledge is, from which we can proceed with some security (p. 4). How much, for example, does our own construction of the social world influence what we regard as the population and the sample?

The point here is to inject a cautionary note: much of the material that follows can be regarded as 'technical' knowledge, but this would be mistaken, as our point here is that behind that technical knowledge reside ontological and epistemological issues – one cannot simply read off a sampling strategy or design mechanistically. There are no 'hard and fast' rules to be followed unthinkingly; rather, decisions on sampling are deliberative, requiring the exercise of judgement and a reflexive attitude to the assumptions that we might all too easily make. Uprichard (2013) makes the point that issues of sample size and sample error, both of which we meet below, are meaningless unless we know how and why they matter at all (p. 7). Researchers, she avers, have to decide when a sample is good enough, or large enough, or small enough, and this is not simply a question of reading off figures from a table, but a deliberative, reflexive, ontological and epistemological matter (p. 7), a matter of praxis in which action and reflection combine. It is problematic. It is in this spirit that we proceed here.

12.2 The sample size

A question that novice researchers often ask is just how large their samples for the research should be. This is a deceptively simple question but there is no clear-cut or simple answer, for the sample size depends on a large array of factors:

- the research purposes, questions and design;
- the size and nature of the population from which the sample is drawn;
- the heterogeneity of the population from which the sample is drawn;
- the confidence level and confidence interval required;
- the level of accuracy required (the smallest sampling error to be tolerated);
- the statistical power required;
- the representativeness of the population sought in the sample;
- the allowances to be made for attrition and non-response;
- the number of strata in the sample;
- the variability of the factor under study;
- the number of variables included in the research;
- the statistics to be used;
- the scales being used;
- the kind(s) of sample to be used;
- the nature of the research (e.g. quantitative, qualitative, mixed methods).

However, it is possible to give some advice on this matter. Generally speaking, for quantitative research, the larger the sample the better, as this not only gives greater reliability but also enables more sophisticated statistics to be used.

Thus, a sample size of thirty is held by many to be the minimum number of cases if researchers plan to use some form of statistical analysis on their data, though this is a very small number and we would advise very considerably more. Researchers need to think, in advance of any data collection, of the sorts of relationships that they wish to explore within sub-groups of their eventual sample. The number of variables researchers set out to control in their analysis, and the types of statistical tests that they wish to make, must inform their decisions about sample size prior to the actual research undertaking. Typically an anticipated minimum of thirty cases per variable should be used as a 'rule of thumb', i.e. one must be assured of having a minimum of thirty cases for each variable (of course, the thirty cases for variable one could also be the same thirty for variable two), though this is a very low estimate indeed. This number rises rapidly if different sub-groups of the population are included in the sample (discussed below), which is frequently the case.

Further, depending on the kind of analysis to be performed, some statistical tests will require larger samples. For example, let us imagine that one wished to calculate the chi-square statistic, a commonly used test (see Part 5) with crosstabulated data, for example, looking at two sub-groups of stakeholders in a primary school containing sixty ten-year-olds and twenty teachers and their responses to a question on a five-point scale:

Variable: Ten-year-olds should do one hour's homework each weekday evening					
	Strongly disagree	*Disagree*	*Neither agree nor disagree*	*Agree*	*Strongly agree*
10-year-old pupils in the school	25	20	3	8	4
Teachers in the school	6	4	2	4	4

Here the sample size is eighty cases, an apparently reasonably sized sample. However, six of the ten cells of responses (60 per cent) contain fewer than five cases. The chi-square statistic requires there to be five cases or more in 80 per cent of the cells (i.e. eight out of the ten cells). In this example only 40 per cent of the cells contained more than five cases, so even with a comparatively large sample, the statistical requirements for reliable data with a straightforward statistic such as chi-square have not been met. The message is clear, one needs to anticipate, as far as one is able, some possible distributions of the data and see if these will prevent appropriate statistical analysis; if the distributions look unlikely to enable reliable statistics to be calculated then one should increase the sample size, or exercise great caution in interpreting the data because of problems of reliability, or not use particular statistics, or, indeed, consider abandoning the exercise if the increase in sample size cannot be achieved.

The point here is that each variable may need to be ensured of a reasonably large sample size. Indeed Gorard (2003, p. 63) suggests that one can start from the minimum number of cases required in each cell, multiply this by the number of cells, and then double the total. In the example above, with six cases in each cell, the minimum sample would be 120 ($6 \times 10 \times 2$), though to be on the safe side, to try to ensure ten cases in each cell a minimum sample of 200 might be better ($10 \times 10 \times 2$), though even this is no guarantee that the distributions will be safe.

The issue arising out of the example here is also that one can observe considerable variation in the responses from the participants in the research. Gorard (2003, p. 62) suggests that if a phenomenon contains a lot of potential variability then this will increase the sample size. Surveying a variable such as IQ, for example, with a potential range from 70 to around 150, may require a larger sample rather than a smaller sample.

As well as the requirement of a minimum number of cases in order to examine relationships between subgroups, researchers must obtain the minimum sample size that will accurately represent the population being targeted. With respect to size, will a large sample guarantee representativeness? Not necessarily! In our first example, the researcher could have interviewed a total sample of 450 females and still not have represented the male population. Will a small size guarantee representativeness? Again, not necessarily! The latter falls into the trap of saying that 50 per cent of those who expressed an opinion said that they enjoyed science, when the 50 per cent was only one student, as the researcher interviewed only two students in all. Too large a sample might become unwieldy and too small a sample might be unrepresentative (e.g. in the first example, the researcher might have wished to interview 450 students but this would have been unworkable in practice or the researcher might have interviewed only ten students, which, in all likelihood, would have been unrepresentative of the total population of 900 students).

Where simple random sampling is used, the sample size needed to reflect the population value of a particular variable depends both on the size of the population and the amount of heterogeneity in the population (Bailey, 1994). Generally, for populations of equal heterogeneity or variance, the larger the population, the larger the sample that must be drawn. For populations of equal size and the greater the heterogeneity on a particular variable, the larger the sample that is needed. If the population is heterogeneous then a large sample is preferable; if the population is homogeneous then a smaller sample is possible. To the extent that a sample fails to represent accurately the population involved, there is sampling error, discussed below.

Sample size is also determined to some extent by the style of the research. For example, a survey style usually requires a large sample, particularly if inferential statistics are to be calculated. In ethnographic or qualitative research it is more likely that the sample size will be small. Sample size might also be constrained by cost – in terms of time, money, stress, administrative support, the number of researchers, and resources.

Borg and Gall (1979, pp. 194–5) suggest that correlational research requires a sample size of no fewer than thirty cases, that causal-comparative and experimental methodologies require a sample size of no fewer than fifteen cases and that survey research should have

no fewer than 100 cases in each major sub-group and twenty to fifty in each minor sub-group. They advise that sample size has to begin with an estimation of the smallest number of cases in the smallest sub-group of the sample, and 'work up' from that, rather than vice versa (p. 186). So, for example, if 5 per cent of the sample must be teenage boys, and this sub-sample must be thirty cases (e.g. for correlational research), then the total sample will be $30 \div 0.05 = 600$; if 15 per cent of the sample must be teenage girls and the sub-sample must be forty-five cases, then the total sample must be $45 \div 0.15 = 300$ cases.

The size of a probability (e.g. random) sample can be determined in two ways, either by the researcher exercising prudence and ensuring that the sample represents the wider features of the population with the minimum number of cases or by using a table which, from a mathematical formula, indicates the appropriate size of a random sample for a given number of the wider population (see Table 12.1). One such example is provided by Krejcie and Morgan (1970), whose work suggests that if the researcher were devising a sample from a wider population of thirty or fewer (e.g. a class of students or a group of young children in a class) then she/he would be well advised to include the whole population as the sample.

Krejcie and Morgan indicate that the smaller the number of cases there are in the population, the larger the proportion of that population must be which appears in the sample. The converse of this is true: the larger the number of cases there are in the population, the smaller the proportion of that population can appear in the sample. They note that as the population increases, the proportion of the population required in the sample diminishes and, indeed, remains constant at around 384 cases (p. 610). Hence, for example, a piece of research involving all the children in a small rural primary or elementary school (up to 100 students in all) might require between 80 per cent and 100 per cent of the school to be included in the sample, whilst a secondary school of 1,200 students might require a sample of 25 per cent of the school in order to achieve randomness. As a rough guide in a random sample, the larger the sample, the greater is its chance of being representative.

In determining sample size for a probability sample, one has to consider not only the population size but also the error margins that one wishes to tolerate. These are expressed in terms of the confidence level and confidence interval. The confidence *level*, usually expressed as a percentage (usually 95 or 99 per cent), is an index of how sure we can be (e.g. 95 per cent of the time or 99 per cent of the time) that the responses lie within a given variation range. The confidence *interval* is that degree of variation or variation range (e.g. ±1 per cent, or ±2 per cent, or ±3 per cent) that one wishes to ensure.

For example, the confidence *interval* in many opinion polls is ±3 per cent; this means that, if a voting survey indicates that a political party has 52 per cent of the votes then it could be as low as 49 per cent (52 − 3) or as high as 55 per cent (52 + 3). The confidence interval is affected by sample size, population size and the percentage of the sample giving the 'true' answer. A confidence *level* of 95 per cent here would indicate that we could be 95 per cent sure that this result will be within this range of 46 to 55, i.e. ±3 per cent. The confidence level is calculated statistically, based on sample size, confidence level and the percentages of an area under the normal curve of distribution, for example, a 95 per cent confidence level covers 95 per cent of the curve of distribution.

If we want to have a very high confidence level (say 99 per cent of the time) then the sample size will be high. On the other hand, if we want a less stringent confidence level (say 90 per cent of the time), then the sample size will be smaller. Usually a compromise is reached, and researchers opt for a 95 per cent confidence level. Similarly, if we want a very small confidence interval (i.e. a limited range of variation, e.g. 3 per cent) then the sample size will be high, and if we are comfortable with a larger degree of variation (e.g. 5 per cent) then the sample size will be lower.

Some research may require a very stringent confidence level and confidence interval (e.g. 99 per cent and 1 per cent respectively) to ensure certainty. For example, medical research, say for a new drug, cannot tolerate errors as the incorrect result could be fatal. Other kinds of research may be content with a less stringent requirement (e.g. 95 per cent confidence level and 3 per cent confidence interval).

A full table of sample sizes for a random probability sample is given in Table 12.1, with three confidence levels (90 per cent, 95 per cent and 99 per cent) and three confidence intervals (5 per cent, 4 per cent and 3 per cent).

Here the size of the sample reduces at an increasing rate as the population size increases; generally (but not always) the larger the population, the smaller the proportion of the probability sample can be. Also, the higher the confidence level, the greater the sample, and the lower the confidence interval, the higher the sample. A conventional sampling strategy will be to use a 95 per cent confidence level and a 3 per cent confidence interval.

There are several websites that offer sample size calculation services for random samples. Some free sites at the time of writing are:

TABLE 12.1 SAMPLE SIZE, CONFIDENCE LEVELS AND CONFIDENCE INTERVALS FOR RANDOM SAMPLES

Population size	*Confidence level 90%*			*Confidence level 95%*			*Confidence level 99%*		
	Confidence interval 5%	*Confidence interval 4%*	*Confidence interval 3%*	*Confidence interval 5%*	*Confidence interval 4%*	*Confidence interval 3%*	*Confidence interval 5%*	*Confidence interval 4%*	*Confidence interval 3%*
30	27	28	29	28	29	29	29	29	30
50	42	45	47	44	46	48	46	48	49
75	59	64	68	63	67	70	67	70	72
100	73	81	88	79	86	91	87	91	95
120	83	94	104	91	100	108	102	108	113
150	97	111	125	108	120	132	122	131	139
200	115	136	158	132	150	168	154	168	180
250	130	157	188	151	176	203	182	201	220
300	143	176	215	168	200	234	207	233	258
350	153	192	239	183	221	264	229	262	294
400	162	206	262	196	240	291	250	289	329
450	170	219	282	207	257	317	268	314	362
500	176	230	301	217	273	340	285	337	393
600	187	249	335	234	300	384	315	380	453
650	192	257	350	241	312	404	328	400	481
700	196	265	364	248	323	423	341	418	507
800	203	278	389	260	343	457	363	452	558
900	209	289	411	269	360	468	382	482	605
1,000	214	298	431	278	375	516	399	509	648
1,100	218	307	448	285	388	542	414	534	689
1,200	222	314	464	291	400	565	427	556	727
1,300	225	321	478	297	411	586	439	577	762
1,400	228	326	491	301	420	606	450	596	796
1,500	230	331	503	306	429	624	460	613	827
2,000	240	351	549	322	462	696	498	683	959
2,500	246	364	581	333	484	749	524	733	1,061
5,000	258	392	657	357	536	879	586	859	1,347
7,500	263	403	687	365	556	934	610	911	1,480
10,000	265	408	703	370	566	964	622	939	1,556
20,000	269	417	729	377	583	1,013	642	986	1,688
30,000	270	419	738	379	588	1,030	649	1,002	1,737
40,000	270	421	742	381	591	1,039	653	1,011	1,762
50,000	271	422	745	381	593	1,045	655	1,016	1,778
100,000	272	424	751	383	597	1,056	659	1,026	1,810
150,000	272	424	752	383	598	1,060	661	1,030	1,821
200,000	272	424	753	383	598	1,061	661	1,031	1,826
250,000	272	425	754	384	599	1,063	662	1,033	1,830
500,000	272	425	755	384	600	1,065	663	1,035	1,837
1,000,000	272	425	756	384	600	1,066	663	1,036	1,840

www.surveysystem.com/sscalc.htm
www.macorr.com/ss_calculator.htm
www.raosoft.com/samplesize.html
www.danielsoper.com/statcalc3/category.aspx?id=19
www.surveymonkey.com/mp/sample-size-calculator

Here the researcher inputs the desired confidence level, confidence interval and the population size, and the sample size is automatically calculated.

A further consideration in the determination of sample size is the kind of variables included. Bartlett *et al*. (2001) indicate that sample sizes for categorical variables (e.g. sex, education level) will differ from those of continuous data (e.g. marks in a test, money in the bank); typically categorical data require larger samples than continuous data. They provide a summary table (Table 12.2) to indicate the different sample sizes required for categorical and continuous data:

Within the discussion of categorical and continuous variables, Bartlett *et al*. (2001, p. 45) suggest that, for categorical data, a 5 per cent margin of error is commonplace, whilst for continuous data, a 3 per cent margin of error is usual, and these are the intervals that they use in their table (Table 12.2). Here, for both categorical and continuous data, the proportion of the population decreases as the sample increases, and, for continuous data, there is no difference in the sample sizes for populations of 2,000 or more. The researcher should normally opt for the larger sample size (i.e. the sample size required for categorical data) if both categorical and continuous data are being used.

Bartlett *et al*. (2001, pp. 48–9) also suggest that the sample size will vary according to the statistics to be used. They suggest that if multiple regressions are to be calculated then 'the ratio of observations [cases] to independent variables should not fall below five', though some statisticians suggest a ratio of 10 : 1, particularly for continuous data, as, in continuous data, the sample sizes tend to be smaller than for categorical data. They also suggest that, in multiple regression: (a) for *continuous* data, if the number of independent variables is in the ratio of 5 : 1 then the sample size should be no fewer than 111 and the number of regressors (independent variables) should be no more than 22; (b) for *continuous* data, if the number of independent variables is in the ratio of 10 : 1 then the sample size should be no fewer than 111 and the number of regressors (independent variables) should be no more than 11; (c) for *categorical* data, if the number of independent variables is in the ratio of 5 : 1 then the sample size should be no fewer than 313 and the number of regressors (independent variables) should be no more than 62; (d) for *categorical* data, if the number of independent variables is

TABLE 12.2 SAMPLE SIZES FOR CATEGORICAL AND CONTINUOUS DATA

Population size	*Sample size*					
	Continuous data (margin of error = 0.3)			*Categorical data (margin of error = 0.05)*		
	alpha = 0.10	*alpha = 0.05*	*alpha = 0.01*	*alpha = 0.10*	*alpha = 0.05*	*alpha = 0.01*
100	46	55	68	74	80	87
200	59	75	102	116	132	154
300	65	85	123	143	169	207
400	69	92	137	162	196	250
500	72	96	147	176	218	286
600	73	100	155	187	235	316
700	75	102	161	196	249	341
800	76	104	166	203	260	363
900	76	105	170	209	270	382
1,000	77	106	173	213	278	399
1,500	79	110	183	230	306	461
2,000	83	112	189	239	323	499
4,000	83	119	198	254	351	570
6,000	83	119	209	259	362	598
8,000	83	119	209	262	367	613
10,000	83	119	209	264	370	623

Source: Bartlett *et al*. (2001, p. 48), reproduced with permission from R. G. Brookshire and J. E. Bartlett

in the ratio of 10 : 1 then the sample size should be no fewer than 313 and the number of regressors (independent variables) should be no more than 31. Bartlett *et al*. (2001, p. 49) also suggest that, for factor analysis, a sample size of no fewer than 100 observations (cases) should be the general rule. However, size can go as low as thirty cases, and the ratio of sample size to number of variables varies from 5 : 1 to 30 : 1 (Tabachnick and Fidell, 2013).

If different sub-groups or strata are to be used then the requirements placed on the total sample also apply to each sub-group, i.e. each stratum (sub-group) becomes a population. Much educational research and sampling concerns itself with strata rather than whole samples, so the issue is significant, and using strata (sub-groups) can rapidly generate the need for a very large sample. If sub-groups are required then the same rules for calculating overall sample size apply to each of the sub-groups. We consider stratified sampling later in this chapter.

Further, determining the size of the sample will also have to take account of non-response, as it may be that non-respondents are not randomly distributed. As Gorard (2013, p. 88) remarks, there may be 'systematic differences' between those who do and do not respond, between those who do and do not take part in a piece of research. Gorard advocates the use of 'sensitivity analysis' (p. 88) to judge the impact of non-respondents, which involves judging (e.g. by calculation) how much difference the non-respondents would need to make to the overall findings for the findings to be false, for example, to reverse the findings.

Next we consider attrition and respondent mortality. Some participants will fail to return questionnaires, leave the research, return incomplete or spoiled questionnaires (e.g. missing out items, putting two ticks in a row of choices instead of only one). Hence it is advisable to overestimate (oversample) rather than to underestimate the size of the sample required, to build in redundancy (Gorard, 2003, p. 60). Unless one has guarantees of access, response and, perhaps, the researcher's own presence at the time of conducting the research (e.g. presence when questionnaires are being completed), then it might be advisable to estimate up to double the size of required sample in order to allow for such loss of clean and complete copies of questionnaires/responses.

Further, with very small sub-groups of populations, it may be necessary to operate a weighted sample – an oversampling – in order to gain any responses at all as, if a regular sample were to be gathered, there would be so few people included as to risk being unrepresentative of the sub-group in question. A weighted sample, in this instance, is where a higher proportion of the sub-group is sampled, and then the results are subsequently scaled down to be fairer in relation to the whole sample.

In some circumstances, meeting the requirements of sample size can be done on an evolutionary basis. For example, let us imagine that you wish to sample 300 teachers, randomly selected. You succeed in gaining positive responses from 250 teachers to, for example, a telephone survey or a questionnaire survey, but you are fifty short of the required number. The matter can be resolved simply by adding another fifty to the random sample, and, if not all of these are successful, then adding some more until the required number is reached.

Borg and Gall (1979, p. 195) suggest that, as a general rule, sample sizes should be large where:

- there are many variables;
- only small differences or small relationships are expected or predicted;
- the sample will be broken down into sub-groups;
- the sample is heterogeneous in terms of the variables under study;
- reliable measures of the dependent variable are unavailable.

Oppenheim (1992, p. 44) adds to this the point that the nature of the scales to be used also exerts an influence on the sample size: the larger the scale, the larger the sample must be. For nominal data the sample sizes may well have to be larger than for interval and ratio data, i.e. a variant of the issue of the number of sub-groups to be addressed, where the greater the number of sub-groups or possible categories, the larger the sample will have to be.

Borg and Gall (1979) set out a formula-driven approach to determining sample size (see also Moser and Kalton, 1977; Ross and Rust, 1997, pp. 427–38), and they also suggest using correlational tables for correlational studies – available in most texts on statistics – as it were 'in reverse' to determine sample size (p. 201), i.e. looking at the significance levels of correlation coefficients and then reading off the sample sizes usually required to demonstrate that level of significance. For example, a correlational significance level of 0.01 would need a sample size of ten if the required coefficient of correlation is 0.65, or a sample size of twenty if the required correlation coefficient is 0.45, and a sample size of 100 if the required correlation coefficient is 0.20. Again, an inverse proportion can be seen – the larger the sample population, the smaller the

required correlation coefficient can be to be deemed significant.

With both qualitative and quantitative data, the essential requirement is that the sample is representative of the population from which it is drawn. In a dissertation concerned with a life history (i.e. n=1), the sample is the population! In a qualitative study of thirty highly able girls of similar socio-economic background following an A-level Biology course, a sample of five or six may suffice the researcher who is prepared to obtain additional corroborative data by way of validation.

Where there is heterogeneity in the population, then a larger sample must be selected on some basis that respects that heterogeneity. Thus, from a staff of sixty secondary school teachers differentiated by gender, age, subject specialism, management or classroom responsibility etc., it would be insufficient to construct a sample consisting of ten female classroom teachers of arts and humanities subjects.

For quantitative data, a precise sample number can be calculated according to the *level of accuracy* and the *level of probability* that the researcher requires in her work. She can then report in her study the rationale and the basis of her research decision (Blalock, 1979). By way of example, suppose a teacher/researcher wishes to sample opinions of an activity (an extra-curricular event) among 1,000 secondary school students. She intends to use a ten-point scale ranging from 0=totally unsatisfactory to 10=absolutely fabulous. She already has data from her own class of thirty students and suspects that the responses of other students will be broadly similar. Her own students rated the activity as follows: mean score=8.27; standard deviation=1.98. In other words, her students were pretty much 'bunched' about a positive appraisal on the ten-point scale. How many of the 1,000 students does she need to sample in order to gain an accurate (i.e. reliable) assessment of what the whole school (n=1,000) thinks of the extra-curricular event?

It all depends on what degree of accuracy and what level of probability she is willing to accept.

A simple calculation from a formula by Blalock (1979, pp. 215–18) shows that:

- if she is happy to be within ±0.5 of a scale point and accurate 19 times out of 20, then she requires a sample of 60 out of the 1,000;
- if she is happy to be within ±0.5 of a scale point and accurate 99 times out of 100, then she requires a sample of 104 out of the 1,000;
- if she is happy to be within ±0.5 of a scale point and accurate 999 times out of 1,000, then she requires a sample of 170 out of the 1,000;
- if she is a perfectionist and wishes to be within ±0.25 of a scale point and accurate 999 times out of 1,000, then she requires a sample of 679 out of the 1,000.

It is clear that sample size is a matter of judgement as well as mathematical precision; even formula-driven approaches make it clear that there are elements of prediction, standard error and human judgement involved in determining sample size.

12.3 Sampling error

If many samples are taken from the same population, it is unlikely that they will all have characteristics identical with each other or with the population; their means will be different. In brief, there will be sampling error (see Cohen and Holliday, 1979, 1996). Sampling error is often taken to be the difference between the sample mean and the population mean. Sampling error is not necessarily the result of mistakes made in sampling procedures. Rather, variations may occur due to the chance selection of different individuals. For example, if we take a large number of samples from the population and measure the mean value of each sample, then the sample means will not be identical. Some will be relatively high, some relatively low, and many will cluster around an average or mean value of the samples. We show this diagrammatically in Figure 12.1.

Why should this occur? We can explain the phenomenon by reference to the Central Limit Theorem which is derived from the laws of probability. This

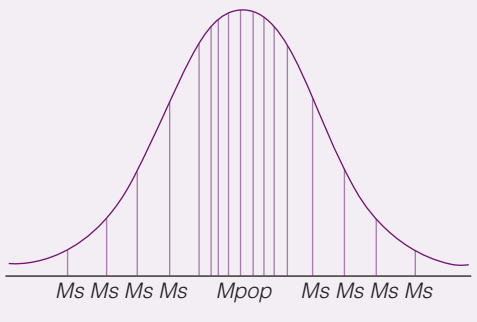


FIGURE 12.1 Distribution of sample means showing the spread of a selection of sample means around the population mean

Source: Cohen and Holliday (1979)

states that if random large samples of equal size are repeatedly drawn from any population, then the mean of those samples will be approximately normally distributed. The distribution of sample means approaches the normal distribution as the size of the sample increases, regardless of the shape – normal or otherwise – of the parent population (Hopkins *et al*., 1996, pp. 159, 388). Moreover, the average or mean of the sample means will be approximately the same as the population mean. The authors demonstrate this (pp. 159–62) by reporting the use of a computer simulation to examine the sampling distribution of means when computed 10,000 times. Rose and Sullivan (1993, p. 144) remind us that 95 per cent of all sample means fall between plus or minus 1.96 standard errors of the sample and population means, i.e. that we have a 95 per cent chance of having a single sample mean within these limits, that the sample mean will fall within the limits of the population mean.

By drawing a large number of samples of equal size from a population, we create a sampling distribution. We can calculate the error involved in such sampling. The standard deviation of the theoretical distribution of sample means is a measure of sampling error (SE) and is called the standard error of the mean (SE_M). Thus,

$$SE = \frac{SD_S}{\sqrt{N}}$$

where SD_S = the standard deviation of the sample and N = the number in the sample.

Strictly speaking, the formula for the standard error of the mean is:

$$SE = \frac{SD_{pop}}{\sqrt{N}}$$

where SD_{pop} = the standard deviation of the population.

However, as we are usually unable to ascertain the *SD* of the total population, the standard deviation of the sample is used instead. The standard error of the mean provides the best estimate of the sampling error. Clearly, the sampling error depends on the variability (i.e. the heterogeneity) in the population as measured by SD_{pop} as well as the sample size (N) (Rose and Sullivan, 1993, p. 143). The smaller the SD_{pop}, the smaller the sampling error; the larger the N, the smaller the sampling error. Where the SD_{pop} is very large, then N needs to be very large to counteract it. Where SD_{pop} is very small, then N, too, can be small and still give a reasonably small sampling error. As the sample size increases, the sampling error decreases. Hopkins *et al*. (1996, p. 159) suggest that, unless there are some very unusual distributions, samples of twenty-five or greater usually yield a normal sampling distribution of the mean; this is comforting!

The standard error of proportions

We said earlier that one consideration in answering 'how big a sample must I obtain?' is 'how accurate do I want my results to be?' This is illustrated in the following example.

A school principal finds that the twenty-five students she talks to at random are reasonably in favour of a proposed change in the lunch break hours, 66 per cent being in favour and 34 per cent being against. How can she be sure that these proportions are truly representative of the whole school of 1,000 students?

A simple calculation of the standard error (SE) of proportions provides the principal with her answer.

$$SE = \frac{P \times Q}{N}$$

where:
P = the percentage in favour
Q = 100 per cent – P
N = the sample size.

The formula assumes that each sample is drawn on a simple random basis. A small correction factor called the finite population correction (*fpc*) is generally applied as follows:

$$\text{SE of proportions} = \sqrt{\frac{(1-f)P \times Q}{N}}$$ where f is the proportion included in the sample.

Where, for example, a sample is 100 out of 1,000, f is 0.1.

$$\text{SE of proportions} = \sqrt{\frac{(1-0.1)(66 \times 34)}{100}} = 4.49$$

With a sample of twenty-five, the SE = 9.4. In other words, the favourable vote can vary between 56.6 per cent and 75.4 per cent; likewise, the unfavourable vote can vary between 43.4 per cent and 24.6 per cent. Clearly, a voting possibility ranging from 56.6 per cent in favour to 43.4 per cent against is less decisive than 66 per cent as opposed to 34 per cent. Should the school principal enlarge her sample to include 100 students, then the SE becomes 4.5 and the variation in the range is reduced to 61.5–70.5 per cent in favour and 29.5–38.5 per cent against. Sampling the whole school's opinion (n = 1,000) reduces the SE to 1.5 and the ranges to 64.5–67.5 per cent in favour and 32.5–35.5 per cent against. It is easy to see why political opinion surveys are often based upon sample sizes of 1,000 to 1,500 (Gardner, 1978).

What is being suggested here generally is that, in order to overcome problems of sampling error, in order to ensure that one can separate random effects and variation from non-random effects, and in order for the power of a statistic to be felt, one should have as large a sample as possible. Samples of fewer than thirty are dangerously small, as they allow the possibility of considerable standard error, and, for over around eighty cases, any increases to the sample size have little effect on the standard error.

12.4 Statistical power and sample size

In calculating sample size, a further consideration is the statistical power required (for quantitative studies), and statistical power influences effect size. We discuss statistical power in Chapter 39; here we refer only to those aspects of statistical power that relate to sample size. Similarly, we mention here the concepts of effect size, statistical significance and one-tailed and two-tailed tests as they relate to sample size, but readers looking for full discussions of these terms should go to Chapter 39.

Statistical power is the probability that a study will detect an effect when there really is an effect there to be detected, separating this from random chance. Power is the probability that a test will correctly reject a false null hypothesis (H_0) and correctly accept the alternative hypothesis (H_1) when it is true, i.e. finding a true effect (see Chapter 39).

Statistical power analysis has four main parameters:

1. The effect size;
2. The sample size (number of observations);
3. The alpha (α) significance level (usually 0.05 or lower);
4. The power of the statistical test (setting the acceptable β level – the probability of committing a Type II error (a false negative) – and the desired power ($1-\beta$), e.g. β of 0.20 and power of 0.80).

Statistical power influences sample size. To calculate the sample size, taking account of statistical power, one needs to set the levels of the alpha (α), the beta (β) and the intended effect size sought (see Chapter 39). Here one can use published tables to determine the sample sizes. Key texts here also include useful guidance and tables, for example, Cohen (1988) and Ellis (2010), setting out sample sizes for different statistical tests (see also Cohen, 1992). Campbell *et al*. (1995) offer useful advice on calculating sample size from power analysis in two-group studies with binary and ordered categorical data (i.e. ordinal data), and they provide tables from which one can read off the sample size required.

Lehr (1992) sets out a straightforward method for calculating sample size needed per group if the power level is 0.80 and the alpha is 0.05, i.e. the two commonly used settings, which is to take the number 16 and divide it by the square of the effect size. Then, for two groups (e.g. a control group and an experimental group) the researcher doubles the result. For example, if the effect size sought is 0.8 (a large effect), then the sample size should be $16/0.8^2 = 16/0.64 = 25$ in each group, 50 in total; if the effect size sought is 0.5 (a moderate effect), then the sample size should be $16/0.5^2 = 16/0.25 = 64$ in each group, 128 in total; if the effect size sought is 0.3 (a small effect), then the sample size should be $16/0.3^2 = 16/0.09 = 177.8$, rounded to 178 in each group, 356 in total. This is an easy-to-use method.

There are many online calculators of sample size which work with effect size, statistical power and the different statistics that the researcher wishes to use, for example:

For a range of statistics: http://powerandsamplesize.com/Calculators
For a range of statistics (go to 'sample size'): www.danielsoper.com/statcalc3
For multiple regression: www.danielsoper.com/statcalc3/calc.aspx?id=1
For hierarchical multiple regression: www.danielsoper.com/statcalc3/calc.aspx?id=16
For t-tests: www.danielsoper.com/statcalc3/calc.aspx?id=47
For post hoc t-tests: www.danielsoper.com/statcalc3/calc.aspx?id=49
For structural equation models: www.danielsoper.com/statcalc3/calc.aspx?id=89
For t-tests and correlations: www.ai-therapy.com/psychology-statistics/sample-size-calculator

In using these sources, both online and in hard copy, the researcher decides the alpha level, the intended effect size (ES) and the statistics to be used (Cohen's *d*, the Pearson correlation, the chi-square, one-way ANOVA, multiple regression, and whether a one-tailed or two-tailed test is being used – see Chapters 40 to 42). From here the researcher can read off the sample size required. For example, setting the power level at 0.80, if one is using Cohen's *d* (a measure of the size of a difference), with an alpha of 0.05 and an effect size of 0.50, and a two-tailed test, then a sample size of 128 people is needed (e.g. 64 in each of two groups between

whom the size of the difference is calculated); with the same alpha and power level, if the effect size is large (0.80) then a sample size of 52 people is needed (e.g. 26 in each of two groups between whom the size of the difference is calculated). For correlations, setting the power level at 0.80, if one is using Pearson's *r* (a measure of association), with an alpha of 0.05 and an effect size (correlation coefficient) of 0.30 then a sample size of 53 people is needed; with the same alpha, if the effect size is large (0.50) then a sample size of 31 people is needed. Examples from these are given in Table 12.3.

Important points to note here are that statistical power and their related sample size calculations vary according to the statistical test used, so researchers must have in mind at the research design stage the statistics that they will use for processing and analysing the numerical data, and they need to decide in advance (Ellis, 2010):

- the type of test to be used, for example, independent t-test, paired t-test, ANOVA, regression etc.;
- the alpha value or significance level to be used (usually 0.01 or 0.05);
- the expected or hoped-for effect size.

Ellis (2010) notes that it is important to know these *before* rather than *after* the data have been collected as it affects decisions on sample size, particularly in the case of small samples (Pituch and Stevens, 2016), though authors note that large samples can often ensure high statistical power. On the other hand, Tabachnick and Fidell (2013) note that there is also a danger of using large samples, as it is almost certain to lead to rejection of the null hypothesis (see Chapter 39).

To improve the statistical power of the test, researchers should strive to use a bigger sample. Torgerson and Torgerson (2008) note that small samples may not be able to detect effect sizes, and that the size of the sample is inversely related to the effect size sought, i.e. if the effect size is expected to be small then a large sample will be needed in order to detect it (p. 128).

Similarly, in using statistical power as part of the calculation of sample size, researchers will need to decide in advance what to set as their alpha (α), beta (β), power levels and their desired effect size. Power analysis is a useful guide to sample size, but caution must be exercised in relying too heavily on it alone, as it is affected by the interaction of several key factors such as effect size, alpha levels and beta levels. Change one of these and the sample size changes. Overall, having as large a sample as possible is desirable for considerations of power analysis and sample size. The work of Ellis (2010) is useful in understanding statistical power and sample size.

12.5 The representativeness of the sample

The researcher will need to consider the extent to which it is important that the sample in fact represents the whole population in question if it is to be a valid sample, to be clear what is being represented, i.e. to set the parameter characteristics of the wider population – the sampling frame – clearly and correctly. There is a popular example of how poor sampling may be unrepresentative and unhelpful for a researcher. A national newspaper reports that one person in every two suffers from backache; this headline stirs alarm in every doctor's surgery throughout the land. However, the newspaper fails to make clear the parameters of the study which gave rise to the headline. It turns out that the research took place (a) in a damp part of the country where the incidence of backache might be expected to be

TABLE 12.3 MINIMUM SAMPLE SIZES AT POWER LEVEL 0.80 WITH TWO-TAILED TEST

	$\alpha = 0.05$			$\alpha = 0.01$		
	Small	*Medium*	*Large*	*Small*	*Medium*	*Large*
	ES = 0.20	ES = 0.50	ES = 0.80	ES = 0.20	ES = 0.50	ES = 0.80
Cohen's *d* (difference test)	788	128	52	1,172	192	78
	ES = 0.10	ES = 0.30	ES = 0.50	ES = 0.10	ES = 0.30	ES = 0.50
Pearson's *r* (measure of association	159	53	31	235	78	45

higher than elsewhere, (b) in a part of the country which contained a disproportionate number of elderly people, again who might be expected to have more backaches than a younger population, (c) in an area of heavy industry where the working population might be expected to have more backache than in an area of lighter industry or service industries, (d) by using two doctors' records only, overlooking the fact that many backache sufferers went to those doctors' surgeries because the two doctors concerned were known to be sympathetic to, rather than responsibly suspicious of, backache sufferers.

These four variables – climate, age group, occupation and reported incidence – exerted a disproportionate effect on the study, i.e. if the study had been carried out in an area where the climate, age group, occupation and reporting were different, then the results might have been different. The newspaper report sensationally generalized beyond the parameters of the data, thereby overlooking the limited representativeness of the study.

It is important to consider adjusting the weightings of sub-groups in the sample once the data have been collected. For example, in a secondary school where half of the students are male and half are female, consider the following table of pupils' responses to the question 'how far does your liking of the form teacher affect your attitude to work?':

Variable: How far does your liking of the form teacher affect your attitude to school work?					
	Very little	*A little*	*Somewhat*	*Quite a lot*	*A very great deal*
Male	10	20	30	25	15
Female	50	80	30	25	15
Total	60	100	60	50	30

Let us say that we are interested in the attitudes according to the gender of the respondents, as well as overall. In this example one could surmise that generally the results indicate that the liking of the form teacher has only a small to moderate effect on the students' attitude to work. However, we have to observe that twice as many girls as boys are included in the sample, and this is an unfair representation of the population of the school, which comprises 50 per cent girls and 50 per cent boys, i.e. girls are over-represented and boys are under-represented. If one equalizes the two sets of scores by gender to be closer to the school population (either by doubling the number of boys or halving the number of girls) then the results look very different, for example:

Variable: How far does your liking of the form teacher affect your attitude to school work?					
	Very little	*A little*	*Somewhat*	*Quite a lot*	*A very great deal*
Male	20	40	60	50	30
Female	50	80	30	25	15
Total	70	120	90	75	45

In this latter case a much more positive picture is painted, indicating that the students regard their liking of the form teacher as a quite important feature in their attitude to school work. Here equalizing the sample to represent more fairly the population by weighting yields a different picture. Weighting the results is important.

12.6 The access to the sample

Access is a key issue and is an early factor that must be decided in research. Researchers will need to ensure not only that access is permitted but is, in fact, practicable. For example, if a researcher were to conduct research into truancy and unauthorized absence from school, and she decided to interview a sample of truants, the research might never commence as the truants, by definition, would not be present! Similarly, access to sensitive areas might be not only difficult but also problematical both legally and administratively, for example, access to child abuse victims, child abusers, disaffected students, drug addicts, school refusers, bullies and victims of bullying. In some sensitive areas access to a sample might be denied by the potential participants themselves, for example, an AIDS counsellor with young people might be so seriously distressed by her work that she simply cannot face discussing with a researcher the subject matter of her traumatic work; it is distressing enough to do the job without living through it again with a researcher.

Access might also be denied by the potential sample participants themselves for very practical reasons, for example, a doctor or a teacher simply might not have the time to spend with the researcher. Further, access might be denied by people who have something to protect, for example, a school which has recently received a very poor inspection result or poor results on external examinations, or a person who has made an important discovery or a new invention and who does not wish to disclose the secret of her success (the trade in intellectual property has rendered this a live issue for many researchers). There are many reasons which

might prevent access to the sample, and researchers cannot afford to neglect this potential source of difficulty in planning research; it is a key issue.

In many cases access is guarded by 'gatekeepers': people who can control the researcher's access to those whom she/he really wants to target. For school staff this might be, for example, headteachers/principals, school governors, school secretaries, form teachers; for students this might be friends, gang members, parents, social workers and so on. It is critical for researchers not only to consider whether access is possible but how access will be sought – to whom does one have to go, both formally and informally, to gain access to the target group.

Not only might access be difficult but its corollary – release of information – might be problematic. For example, a researcher might gain access to a wealth of sensitive information and appropriate people, but there might be a restriction on the release of the data collected; reports may be suppressed, delayed or 'doctored'. It is not always enough to be able to 'get to' the sample, the problem might be to 'get the information out' to the wider public, particularly if it could be critical of powerful people.

12.7 The sampling strategy to be used

There are two main methods of sampling (Cohen and Holliday, 1979, 1982, 1996). The researcher must decide whether to opt for a probability (also known as a random sample) or a non-probability sample (also known as a purposive sample). The difference between them is this: in a probability sample the chances of members of the wider population being selected for the sample are known, whereas in a non-probability sample the chances of members of the wider population being selected for the sample are unknown. In the former (probability sample) every member of the wider population has an equal chance of being included in the sample; inclusion or exclusion from the sample is a matter of chance and nothing else. In the latter (non-probability sample) some members of the wider population definitely will be excluded and others definitely included, i.e. every member of the wider population does not have an equal chance of being included in the sample. In this latter type the researcher has deliberately – purposely – selected a particular section of the wider population to include in or exclude from the sample.

12.8 Probability samples

A probability sample, because it draws randomly from the wider population, is useful if the researcher wishes to be able to make generalizations, because it seeks representativeness of the wider population. (It also permits many statistical tests to be conducted with quantitative data.) This is a form of sampling used in randomized controlled trials. Randomization has two stages – random selection from a population and random allocation to groups (e.g. a control and an experimental group) – and these are key requirements for many experiments and statistics. Randomization, as one of its founding figures, Ronald Fisher (1966), remarked, is designed to overcome myriad within-group and between-group differences. It ensures that the average result, taking into account range and spread, within one group is similar to the average within another group (Torgerson and Torgerson, 2008, p. 29); as the authors remark, '[t]he presence of *all* variables that could affect outcome … in *all* groups will cancel out their effect in the analysis' (p. 29), and if, by chance, other variables are not the same in both groups, then this is unlikely to affect the outcome. Indeed Fisher commented that randomization, intended to overcome individual differences, is sufficient 'to guarantee the validity of the test of significance' in an experiment (1966, p. 21). Randomization has the potential to address external validity, i.e. generalizability, and internal validity, i.e. to avoid selection bias (p. 29).

On the other hand, a non-probability sample deliberately avoids representing the wider population; it seeks only to represent a particular group, a particular named section of the wider population, for example, a class of students, a group of students who are taking a particular examination, a group of teachers.

A probability sample will have less risk of bias than a non-probability sample, whereas, by contrast, a non-probability sample, being unrepresentative of the whole population, may demonstrate skewness or bias. This is not to say that the former is bias-free; there is still likely to be sampling error in a probability sample (discussed below), a feature that has to be acknowledged, for example, opinion polls usually declare their error factors (e.g. ±3%).

There are several types of probability samples: simple random samples; systematic samples; stratified samples; cluster samples; stage samples; and multiphase samples. They all have a measure of randomness built into them and therefore have a degree of generalizability.

Simple random sampling

In simple random sampling, each member of the population under study has an equal chance of being selected and the probability of a member of the population being selected is unaffected by the selection of other members of the population, i.e. each selection is entirely independent of the next. The method involves selecting at random from a list of the population (a sampling frame) the required number of subjects for the sample. This can be done by drawing names out of a hat until the required number is reached, or by using a table of random numbers set out in matrix form (these are reproduced in many books and websites on quantitative research methods and statistics). Researchers can also use software (e.g. SPSS, Excel) to generate random samples and randomly allocate individuals to groups, though some of these might have technical bias in their programming.

Using computer-generated samples for random allocation to different groups may, inadvertently, lead to an imbalance between those groups on key variables of interest (Torgerson, and Torgerson, 2008, p. 31). For example, Garcia *et al*. (2014) encountered this problem in their initial random allocation into two groups (control and experimental) in a school project, which led to imbalance in terms of assessed student performance levels in key subjects, and the random allocation had to be iterated more than once in order to arrive at random allocation which overcame such imbalance. In such cases 'matched randomization' might be considered. Here, for example, a pair of children might be matched on the variables of interest and then one from each pair is randomly allocated to either the control or experimental group (cf. Torgerson and Torgerson, 2008, p. 35).

Addressing probability and chance, the sample should contain subjects with characteristics similar to the population as a whole; some old, some young, some tall, some short, some fit, some unfit, some rich, some poor etc. One potential problem associated with this particular sampling method is that a complete list of the population is needed and this is not always readily available. On the other hand, Table 12.1 indicates the number of people needed in a random sample with regard to the population size, regardless of detailed characteristics of the sample. This requires the researcher to define carefully the population from which the sample is drawn: for example, it is little help in trying to generalize to all the males and females in a school if only males are taken as the population from which the sample is drawn.

Systematic sampling

This method is a modified form of simple random sampling. It involves selecting subjects from a population list in a systematic rather than a random fashion. For example, if from a population of, say, 2,000 a sample of 100 is required, then every twentieth person can be selected. The starting point for the selection is chosen at random.

One can decide how frequently to make systematic sampling by a simple statistic – the total number of the wider population being represented divided by the sample size required:

$$f = \frac{N}{sn}$$

f = frequency interval
N = the total number of the wider population
sn = the required number in the sample.

Let us say that the researcher is working with a school of 1,400 students; by looking at the table of sample size (Table 12.1) required for a random sample of these 1,400 students, she sees that 301 students are required to be in the sample. Hence the frequency interval (*f)* is:

$$\frac{1400}{302} = 4.635 \text{ (which rounds up to 5.0)}$$

Hence the researcher would pick out every fifth name on the list of cases.

Such a process, of course, assumes that the names on the list themselves have been listed in a random order. A list of females and males might list all the females first, before listing all the males; if there were 200 females on the list, the researcher might have reached the desired sample size before reaching that stage of the list which contained males, thereby distorting (skewing) the sample. Another example is where the researcher decides to select every thirtieth person from a list of school students, but it happens that: (a) the school has just over thirty students in each class; (b) each class is listed from high-ability to low-ability students; (c) the school listing identifies the students by class. Here, although the sample is drawn from each class, it is not fairly representing the whole school population since it is drawing almost exclusively on the lower-ability students. This is the issue of *periodicity* (Calder, 1979).

Not only is there the question of the order in which names are listed in systematic sampling, but there is also the issue that this process may violate one of the fundamental premises of probability

sampling, namely that every person has an equal chance of being included in the sample. In the example above where every fifth name is selected, this guarantees that names 1–4, 6–9, etc. will be excluded, i.e. everybody does not have an equal chance to be chosen. The ways to reduce this problem are to ensure that the initial listing is selected randomly and that the starting point for systematic sampling is similarly selected randomly.

Random stratified sampling

Random stratified sampling involves dividing the population into homogeneous groups, each group containing subjects with similar characteristics, and then randomly sampling within those groups. For example, group A might contain males and group B, females. In order to obtain a sample representative of the whole population in terms of sex, a random selection of subjects from group A and group B must be taken. If needed, the exact proportion of males to females in the whole population can be reflected in the sample. For example, if a school has a population with 75 per cent of students whose first language is English and 25 per cent with a different first language then the researcher can randomly sample to contain 75 per cent of first-language English speakers and 25 per cent with different first languages, in order to keep the proportions in the sample the same as those in the population. The researcher will need to identify those characteristics of the wider population which must be included in the sample, i.e. to identify the parameters of the wider population. This is the essence of establishing the sampling frame.

To organize a stratified random sample is a simple two-stage process. First, identify those characteristics that appear in the wider population which must also appear in the sample, i.e. divide the wider population into homogeneous and, if possible, discrete groups (strata), for example, males and females. Second, randomly sample within these groups, the size of each group being determined either by the judgement of the researcher or by reference to Tables 12.1 or 12.2.

The decision on which characteristics to include should strive for simplicity as far as possible, as the more factors there are, not only the more complicated the sampling becomes, but often the larger the sample will have to be in order to include representatives of all strata of the wider population. For example, imagine that we are surveying a whole school of 1,000 students in a multi-ethnic school. Table 12.1 suggests that we need 278 students in our random sample, to ensure representativeness. However, let us imagine that we wished to stratify our groups into, for example, Chinese (50 students), Spanish (100 students), English (800 students) and Arabic (50 students). From tables of random sample sizes we work out a random sample *with* stratification, i.e. for each stratum, which yields the following:

Students	*Population*	*Sample*
English-speakers	800	260
Spanish-speakers	100	80
Arabic-speakers	50	44
Mandarin-speakers	50	44
Total	1,000	428

Our original sample size of 278 has now increased, very quickly, to 428. The message is very clear: the more strata (sub-groups) we have, the larger the sample will be. Hence the advice here is to have as few strata as is necessary, but no fewer.

A random stratified sample is a useful blend of randomization and categorization, thereby enabling both a quantitative and qualitative piece of research to be undertaken. Quantitative research can use statistical analysis, whilst qualitative research can target those groups in institutions or clusters of participants who might be approached to participate in the research.

Cluster sampling

When the population is large and widely dispersed, gathering a simple random sample poses administrative problems. Suppose we want to survey students' fitness levels in a particularly large community or across a country. It would be completely impractical to select students randomly and spend an inordinate amount of time travelling about in order to test them. By cluster sampling, the researcher can select a specific number of schools and test all the students in those selected schools, i.e. a geographically close cluster is sampled.

One has to be careful to ensure that cluster sampling does not build in bias. For example, let us imagine that we take a cluster sample of a city in an area of heavy industry or great poverty; this may not represent all kinds of cities or socio-economic groups, i.e. there may be similarities within the sample that do not catch the variability of the wider population. The issue here is one of representativeness; hence it might be safer to take several clusters and to sample lightly within each cluster, rather than to take fewer clusters and sample heavily within each.

Cluster samples are widely used in small-scale research. In a cluster sample the parameters of the wider population are often drawn very sharply; a researcher, therefore, would have to comment on the generalizability of the findings. The researcher may also need to stratify within this cluster sample if useful data, i.e. those which are focused and which demonstrate discriminability, are to be acquired.

Stage sampling

Stage sampling is an extension of cluster sampling. It involves selecting the sample in stages, that is, taking samples from samples. For example, one type of stage sampling might be to select a number of schools at random, and from within each of these schools, select a number of classes at random, and from within those classes select a number of students.

Morrison (1993, pp. 121–2) provides an example of stage sampling. Let us say that a researcher wants to administer a questionnaire to all sixteen-year-olds in secondary schools in one region, and chooses eleven such schools from a population of, say, fifteen schools. By contacting the eleven schools she finds that there are 2,000 sixteen-year-olds on roll. Because of questions of confidentiality she is unable to find out the names of all the students so it is impossible to draw their names out of a hat to achieve randomness (and even if she had the names, it would be a mind-numbing activity to write out 2,000 names to draw out of a hat!). From looking at Table 8.1 she finds that, for a random sample of the 2,000 students, the sample size is 322 students. How can she proceed?

The first stage is to list the eleven schools on a piece of paper and then to put the names of the eleven schools onto a small card and place each card in a hat. She draws out the first name of the school, puts a tally mark by the appropriate school on her list and returns the card to the hat. The process is repeated 321 times, bringing the total to 322. The final totals might appear thus:

School	1	2	3	4	5	6	7	8	9	10	11	Total
Required number of students	22	31	32	24	29	20	35	28	32	38	31	322

For the second stage she then approaches the eleven schools and asks each of them to select randomly the required number of students for each school. Randomness has been maintained in two stages and a large number (2,000) has been rendered manageable. The process at work here is to go from the general to the specific, the wide to the focused, the large to the small. Caution has to be exercised here, as the assumption is that the schools are of the same size and are large; that may not be the case in practice, in which case this strategy may be inadvisable.

The issue can become more complex, as the eleven schools are a sample of the population of the schools in the region, raising the question of what the sample is: the eleven schools or the 322 students (cf. Gorard, 2013, pp. 82–3). Whilst the eleven schools are the random sample from the population of fifteen schools, the 322 students are a clustered sample from the eleven schools. Gorard provides some useful advice here: if the intention is to *compare* institutions then the sample size here would be eleven, and if the intention is to look at overall results then the sample size is 322.

Multi-phase sampling

In stage sampling there is a single unifying purpose throughout the sampling. In the previous example the purpose was to reach a particular group of students from a particular region. However, in a multi-phase sample the purposes change at each phase, for example, at phase one the selection of the sample might be based on the criterion of geography (e.g. students living in a particular region); phase two might be based on an economic criterion (e.g. schools whose budgets are administered in markedly different ways); phase three might be based on a political criterion (e.g. schools whose students are drawn from areas with a tradition of support for a particular political party), and so on. Here the sample population changes at each phase of the research.

12.9 Non-probability samples

The selectivity which is built into a non-probability sample derives from the researcher targeting a particular group, in the full knowledge that it does not represent the wider population; it simply represents itself. This is frequently the case in small samples or small-scale research, for example, one or two schools, two or three groups of students, a particular group of teachers, where no attempt to generalize is desired. It is also frequently the case for ethnographic research, action research or case study research. Small-scale research often uses non-probability samples because, despite their non-representativeness, they are far less complicated to set up, are considerably less expensive and can prove perfectly adequate where researchers do not seek to generalize their findings beyond the sample in question, or where they are simply piloting a questionnaire as a prelude to the main study.

Just as there are several types of probability sample, so there are several types of non-probability sample: convenience sampling, quota sampling, dimensional sampling, purposive sampling and snowball sampling. Each type of sample seeks only to represent itself or instances of itself in a similar population, rather than attempting to represent the whole, undifferentiated population.

Convenience sampling

Convenience sampling, or, as it is sometimes called, accidental or opportunity sampling, involves choosing the nearest individuals to serve as respondents and continuing that process until the required sample size has been obtained of those who happen to be available and accessible at the time. Captive audiences such as students or student teachers often serve as respondents based on convenience sampling. The researcher simply chooses the sample from those to whom she has easy access. As it does not represent any group apart from itself, it does not seek to generalize to the wider population; for a convenience sample that is an irrelevance. The researcher, of course, must take pains to report this point – that generalizability in this type of sampling is negligible. A convenience sample may be selected for a case study or a series of case studies.

Quota sampling

Quota sampling has been described as the non-probability equivalent of stratified sampling (Bailey, 1994). Like a stratified sample, a quota sample strives to represent significant characteristics (strata) of the wider population and sets out to represent these in the proportions in which they can be found in the wider population. For example, suppose that the wider population comprised 55 per cent females and 45 per cent males, then the sample would have to contain 55 per cent females and 45 per cent males; if the population of a school contained 80 per cent of students up to and including the age of sixteen, and 20 per cent of students aged seventeen and over, then the sample would have to contain 80 per cent of students up to the age of sixteen and 20 per cent of students aged seventeen and above. A quota sample, then, seeks to give proportional weighting to selected factors (strata) which reflects their weighting/proportions in the wider population. The researcher wishing to devise a quota sample can proceed in three stages:

Stage 1: Identify those characteristics (factors) that appear in the wider population which must also appear in the sample, i.e. divide the wider population into homogeneous and, if possible, discrete groups (strata), for example, males and females, Asian, Chinese and African-Caribbean.

Stage 2: Identify the proportions in which the selected characteristics appear in the wider population, expressed as a percentage.

Stage 3: Ensure that the percentaged proportions of the characteristics selected from the wider population appear in the sample.

Ensuring correct proportions in the sample may be difficult to achieve if the proportions in the wider community are unknown or if access to the sample is difficult; sometimes a pilot survey might be necessary in order to establish those proportions (and even then sampling error or a poor response rate might render the pilot data problematical).

It is straightforward to determine the minimum number required in a quota sample. Let us say that the total number of students in a school is 1,700, comprising:

Performing arts	300 students
Natural sciences	300 students
Humanities	600 students
Business and social sciences	500 students

The proportions being 3 : 3 : 6 : 5, a minimum of seventeen students might be required (3 + 3 + 6 + 5) for the sample. Of course, this would be a minimum only, and it might be desirable to go higher than this. The price of having too many characteristics (strata) in quota sampling is that the minimum number in the sample very rapidly can become very large, hence in quota sampling it is advisable to keep the numbers of strata to a minimum. The larger the number of strata, the larger the number in the sample will become, often very quickly.

Purposive sampling

In purposive sampling, often (but by no means exclusively) a feature of qualitative research, researchers handpick the cases to be included in the sample on the basis of their judgement of their typicality or possession of the particular characteristic(s) being sought. They assemble the sample to meet their specific needs.

Purposive sampling is undertaken for several kinds of research (Teddlie and Yu, 2007), including: to achieve representativeness, to enable comparisons to be made, to focus on specific, unique issues or cases, to generate theory through the gradual accumulation of data from different sources. Purposive sampling, Teddlie and Yu aver, involves a trade-off: on the one hand it provides greater depth to the study than probability sampling; on the other hand it provides less breadth to the study than probability sampling.

As its name suggests, a purposive sample has been chosen for a specific purpose, for example: (a) a group of principals and senior managers of secondary schools is chosen as the research is studying the incidence of stress among senior managers; (b) a group of disaffected students has been chosen because they might indicate most distinctly the factors which contribute to students' disaffection (they are *critical cases*, akin to 'critical events' discussed in Chapter 33, or *deviant cases* – those cases which go against the norm) (Anderson and Arsenault, 1998, p. 124); (c) one class of students has been selected to be tracked throughout a week in order to report on the curricular and pedagogic diet offered to them so that other teachers in the school can compare their own teaching to that reported. Whilst this type of sample may satisfy the researcher's needs, it does not pretend to represent the wider population; it is deliberately and unashamedly selective and biased.

In many cases purposive sampling is used in order to access 'knowledgeable people', i.e. those who have in-depth knowledge about particular issues, maybe by virtue of their professional role, power, access to networks, expertise or experience (Ball, 1990). There is little benefit in seeking a random sample when most of the random sample may be largely ignorant of particular issues and unable to comment on matters of interest to the researcher, in which case a purposive sample is vital. Though they may not be representative and their comments may not be generalizable, this is not the primary concern in such sampling; rather the concern is to acquire in-depth information from those who are in a position to give it.

Another variant of purposive sampling is the *boosted* sample. Gorard (2003, p. 71) comments on the need to use a boosted sample in order to include those who may otherwise be excluded from, or under-represented in, a sample because there are so few of them. For example, one might have a very small number of special needs teachers or students in a primary school or nursery, or one might have a very small number of children from certain ethnic minorities in a school, such that they may not feature in a sample. In this case the researcher will deliberately seek to include a sufficient number of them to ensure appropriate statistical analysis or representation in the sample, adjusting any results from them, through weighting, to ensure that they are not over-represented in the final results. This is an endeavour, perhaps, to meet the demands of social inclusion.

A further variant of purposive sample is *negative case sampling*. Here the researcher deliberately seeks those people who might disconfirm the theories being advanced (the Popperian equivalent of falsifiability), thereby strengthening the theory if it survives such potentially disconfirming cases. A softer version of negative case sampling is *maximum variation sampling*, selecting cases which are as varied as possible on the issue in question (Anderson and Arsenault, 1998, p. 124) in order to ensure strength and richness to the data, their applicability and their interpretation. In this latter case, it is almost inevitable that the sample size will increase or be large.

Teddlie and Yu (2007), Teddlie and Tashakkori (2009, p. 174) and Flick (2009, pp. 122–3) provide a typology of several kinds of purposive sample; they group these under several main areas. In terms of sampling, in order to achieve representativeness or comparability they include several types of purposive sample:

- typical case sampling, in which the sample includes the most typical cases of the group or population under study, i.e. representativeness;
- extreme or deviant case sampling, in which the most extreme cases (at either end of a continuum, e.g. success and failure, tolerance and intolerance, most and least stressed) are studied in order to provide the most outstanding examples of a particular issue, to compare with the typical cases (i.e. comparability) or to expose issues that might not otherwise present themselves (e.g. what can happen when a young child is exposed to drug pushers, family violence or repeated failure at school);
- intensity sampling of a particular group (e.g. highly effective teachers, highly talented children) in which the sample provides clear examples of the issue in question;
- maximum variation sampling, in which samples are chosen that possess or exhibit a very wide range of characteristics or behaviours respectively in connection with a particular issue;
- homogeneous sampling, in which the samples are chosen for their similarity (which can then be used for contrastive analysis or comparison with maximum variation groups or intensity sampling of other groups);
- reputational case sampling, in which samples are selected by key informants, on the recommendation of others or because the researchers are aware of their characteristics (e.g. a Minister of Education, a politician) – see below, snowball sampling and respondent-driven sampling;
- criterion sampling, in which all the cases are sampled which fit a particular criterion being studied.

In terms of sampling of special or unique cases, purposive sampling includes four types:

- revelatory case sampling, in which individuals are approached because they are members of a particular group and can reveal heretofore unknown insights, for example, fundamentalist religious schools, schools for refugees or single-ethnic minorities;
- critical case sampling: a widely used sampling technique, akin to extreme case sampling, in which a particular individual, group of individuals or cases is studied in order to yield insights that might have wider application, for example, Tripp's (1993) study of critical incidents in teaching, or Morrison's (2006) study of sensitive educational research, focusing on small states and territories, which treats one small territory as a critical case study of issues in the fields in question, which are felt to be their strongest, and which can illuminate issues in the topic which are of wider concern for other similar small states and territories;
- politically important case sampling, for example, Ball's (1990) interviews with senior politicians and Bowe's *et al.*'s (1992) interviews with a UK cabinet minister and politicians;
- complete collection sampling, in which all the members of a particular group are included, for example, all the high-achieving, musically gifted students in a sixth form.

Teddlie and Tashakkori (2009, p. 174) also indicate four examples of 'sequential sampling' in their typologies of purposive sampling:

- theoretical sampling (discussed below, cf. Glaser and Strauss, 1967), in which those cases are selected that will yield greater insight into the theoretical issue(s) under investigation. As Glaser and Strauss (1967, p. 45) suggest, the data collection is for theory generation, and, as the theory emerges, so will the next step in the data collection suggest itself, i.e. the theory drives the investigation. An example of this might be in examining childhood poverty, in which the researchers might look at those who have always been poor, those who have moved out of – or into – poverty, rural poverty, urban poverty, poverty in small families, poverty in large families, poverty in single parent families, and so on;
- conforming and disconforming case sampling, in which samples are selected from those that do and do not conform to typical trends or patterns, in order to study the causes or reasons for their conformity or disconformity;
- opportunistic sampling (see also above, convenience sampling), in which further individuals or groups are sampled as the research develops or changes and which, as validity and reliability dictate, should be included;
- snowball sampling (discussed below), in which researchers use social networks, informants and contacts to put them in touch with further individuals or groups.

Purposive sampling is a key feature of qualitative research.

Dimensional sampling

One way of reducing the problem of sample size in quota sampling is to opt for dimensional sampling. Dimensional sampling, a refinement of quota sampling, involves identifying various factors of interest in a population and obtaining at least one respondent of every combination of those factors. Thus, in a study of racism, for example, researchers may wish to distinguish first-, second- and third-generation immigrants. Their sampling plan might take the form of a multi-dimensional table with 'ethnic group' across the top and 'generation' down the side. A second example might be of a researcher who may be interested in studying disaffected students, girls and secondary-aged students and who may find a single disaffected secondary female student, i.e. a respondent who is the bearer of all of the characteristics sought.

Snowball sampling

In snowball sampling researchers identify a small number of individuals who have the characteristics in which they are interested. These people are then used as informants to identify, or put the researchers in touch with, others who qualify for inclusion; these, in turn, identify yet others – hence the term snowball sampling (also known as 'chain-referral methods'). This method is useful for sampling a population where access is difficult, maybe because the topic for research (and hence the sample) is sensitive (e.g. teenage solvent abusers; issues of sexuality; criminal gangs), or where participants might be suspicious of researchers, or where contact is difficult, for example, those without telephones, the homeless (Heckathorn, 2002). As Faugier and Sargeant (1997), Browne (2005) and Morrison (2006) argue, the more sensitive the research, the more difficulty there is in sampling and gaining access to a sample.

Hard-to-reach groups include minorities, marginalized or stigmatized groups, 'hidden groups' (those who

do not wish to be contacted or reached, e.g. drug pushers, gang members, sex workers, problem drinkers or gamblers, residents of 'safe houses' or women's refuges), old or young people with disabilities, the very powerful or social elite (Noy, 2008), dispersed communities (e.g. rural farm workers) (Brackertz, 2007).

Snowball sampling is also useful where communication networks are undeveloped (e.g. where a researcher wishes to interview stand-in teachers – teachers who are brought in on an ad hoc basis to cover for absent regular members of a school's teaching staff – but finds it difficult to acquire a list of these stand-in teachers), or where an outside researcher has difficulty in gaining access to schools (going through informal networks of friends/acquaintance and their friends and acquaintances and so on rather than through formal channels). The task for the researcher is to establish who are the critical or key informants with whom initial contact must be made.

Snowball sampling is particularly valuable in qualitative research, indeed is often pre-eminent in qualitative research; it is a means in itself, rather than a default, fall-back position (Noy, 2008, p. 330). It uses participants' social networks and personal contacts for gaining access to people. In snowball sampling, interpersonal relations feature very highly (Browne, 2005), as the researcher is reliant on: (a) friends, friends of friends, friends of friends of friends; (b) acquaintances, acquaintances of acquaintances, acquaintances of acquaintances of acquaintances; (c) contacts (personally known or not personally known), contacts of contacts, contacts of contacts of contacts. 'Snowball sampling is essentially social' (Noy, 2008, p. 332), as it often relies on strong interpersonal relations, known contacts and friends; it requires social knowledge and an equalization of power relations (Noy, 2008, p. 329). In this respect it reduces or even dissolves asymmetrical power relations between researcher and participants, as the contacts might be built on friendships, peer group membership and personal contacts and because participants can act as gatekeepers to other participants and informants exercise control over whom else to involve and refer. Indeed in respondent-driven sampling (discussed below), a variant of snowball sampling, the respondents not only identify further contacts for the researcher but actively recruit them to be involved in the research (Heckathorn, 1997, p. 178), i.e. participants who might be initially uncooperative with researchers might be cooperative for their peer group members who approach them (p. 197). Snowball sampling here, then, is 'respondent driven' (Heckathorn, 1997, 2002), where respondents identify others for the researcher to contact.

In researching 'hidden populations' typically there are no sampling frames, so researchers do not know the population from which the sample can be drawn, and there is often a problem of access as such groups may guard their privacy (e.g. if their behaviour is illegal, or stigmatized) and, even if access is gained, truthful responses may not be forthcoming as participants may deliberately conceal the truth in order to protect themselves (Heckathorn, 1997, p. 174).

Snowball sampling may rely on personal, social contacts, but it can also rely on 'reputational contacts' (e.g. Farquharson, 2005), where people may be able to identify to the researcher other known persons in the field. The 'reputational snowball' (p. 347) can be a powerful means of identifying significant others in a 'micro-network' (p. 349), particularly if one is researching powerful individuals and policy makers who are not always known to the public. As Farquharson (2005, p. 346) remarks, 'policy networks' are groups of interconnected institutions and/or people who are influential in the field, perhaps to advance, promote, block, develop or initiate policy. A reputational snowball can be generated by asking individuals – either at interview or by open-ended questions on a questionnaire – to identify others in the field who are particularly influential, important or worth contacting.

On the one hand, snowball sampling can reach the hard-to-reach, not least if the researcher is a member of the groups being researched (e.g. Browne's (2005) study of non-heterosexual women, of which she was one and therefore had her own circle of friends and contacts, and in which rapport and trust were easier to establish).

On the other hand, snowball sampling can be prone to biases stemming from the influence of the initial contact and the problem of volunteer-only samples (Heckathorn, 2002, p. 12). Browne (2005) indicates that, because she was a member of a white, middle-class group of non-heterosexual women, her contacts tended to be from similar backgrounds, and other non-heterosexual women were not included because they were not in the same 'loop' of social contacts. In other words, snowball sampling is influenced by the researcher's initial points of contact, as these drive the subsequent contacts, and, indeed, can lead to sampling or over-sampling of cooperative groups or individuals (Heckathorn, 1997, p. 175). Two methods can be employed to overcome this: (a) key informant sampling asks participants about others' behaviours (but this raises the problem of informed consent and confidentiality of others) (Heckathorn, 2002, p. 13), whilst (b) targeted sampling tries to ensure a non-biased sample, to include all those who should be included (i.e. to

prevent under-sampling) and who represent different facets of the issue or group under study (see Heckathorn (1997, 2002) for a fuller discussion of this matter and for how to address and overcome bias in respondent-driven samples).

Further, if a researcher is to move beyond his or her personal contacts, to try to be more inclusive of otherwise excluded sub-groups or individuals, then there is a risk in having such small numbers of others that tokenism is at work. Browne (2005, p. 53) writes that the women who participated in her research were also gatekeepers of contact to other non-heterosexual women who, for a variety of reasons (not least of which was the wish to avoid revealing too much to a friend), may not have wished to be involved. Bias can both include and exclude members of a population and a sample; it 'can create other "hidden populations"' (Browne, 2005, p. 53), and the gatekeepers can protect friends by not referring them to the researcher (Heckathorn, 1997, p. 175).

Figure 12.2 indicates a linear, sequential method of sampling (with unidirectional arrows). Noy (2008, p. 333) comments that, as the ordinal succession proceeds, the later members of the sample might have different characteristics or attributes from the earlier members of the sample, i.e. the sample is not necessarily homogeneous. This is important, as it overcomes the problem indicated earlier, where the influence of initial contacts on later contacts is high; having many waves of contacts reduces this influence (Heckathorn, 1997, p. 197).

Snowball sampling can be used as the main method of gaining access to people or as an auxiliary method of gaining access to people for further, in-depth data collection and exploration of issues.

Volunteer sampling

In cases where access is difficult, the researcher may have to rely on volunteers, for example, personal friends, or friends of friends, or participants who reply to a newspaper advertisement, or those who happen to be interested from a particular school, or those attending courses. Sometimes this is inevitable (Morrison, 2006) as it is the only kind of sampling that is possible, and it may be better to have this kind of sampling than no research at all.

In these cases one has to be very cautious in making any claims for generalizability or representativeness, as volunteers may have a range of different motives for volunteering, for example, wanting to help a friend, interest in the research, wanting to benefit society, an opportunity for revenge on a particular school or headteacher/principal. Volunteers may be well intentioned, but they do not necessarily represent the wider population, and this has to be made clear.

Theoretical sampling

Theoretical sampling is a feature of grounded theory (see Chapter 37). In grounded theory the sample size is relatively immaterial, as one works with the data that one has. Indeed grounded theory would argue that the sample size could be infinitely large, or, as a fall-back

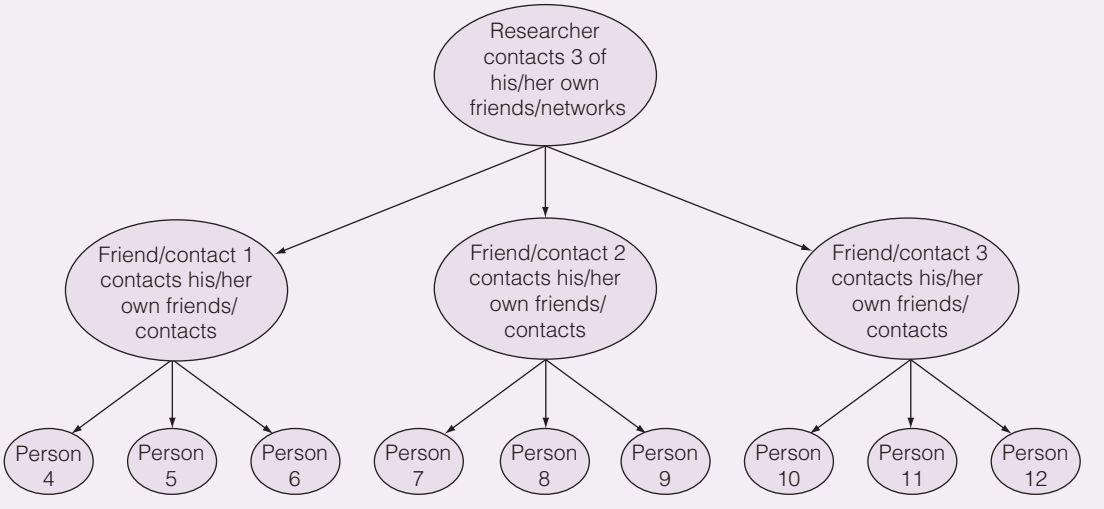


FIGURE 12.2 Snowball sampling

position, large enough to 'saturate' the categories and issues, such that new data do not cause any modification to the theory which has been generated.

Theoretical sampling requires the researcher to have sufficient data to be able to generate and 'ground' the theory in the research context, however defined, i.e. to create a theoretical explanation of what is happening in the situation, without finding any more data that do not fit the theory. Since the researcher will not know in advance how much or what range of data will be required, it is difficult, to the point of impossibility, exhaustion or time limitations, to know in advance the sample size required. Having conducted analysis of collected data, the researcher decides what further data to collect and from whom, in order to develop the emergent theory (Glaser and Strauss, 1967, p. 4). Theoretical sampling places the development of theory as the prime concern (cf. Creswell, 2012, p. 433), and so the researcher gathers more and more data until the theory remains unchanged or until the boundaries of the context of the study have been reached, until no modifications to the grounded theory are made in light of constant comparisons, and this may mean several rounds of data collection from different samples (Flick, 2009, p. 118). 'Theoretical saturation' (Glaser and Strauss, 1967, p. 61) occurs when no additional data are found which advance, modify, qualify, challenge, extend or add to the theory developed (see also Krueger and Casey, 2000).

Two key questions for the grounded theorist using theoretical sampling (Glaser and Strauss, 1967) are: (a) to which groups does one turn next for data? (b) for what theoretical purposes does one seek further data? In response to (a), Glaser and Strauss (p. 49) suggest that the decision is based on theoretical relevance, i.e. those groups that will assist in the generation of as many properties and categories as possible. The size of the data set may be fixed by the number of participants in the organization, or the number of people to whom one has access, but the researcher has to consider that the door may have to be left open for him/her to seek further data in order to ensure theoretical adequacy and to check what has been found so far with further data (Flick *et al*., 2004a, p. 170). In this case it is not always possible to predict at the start of the research just how many, and who, the researcher will need for the sampling; it becomes an iterative process. Flick (2009, p. 118) makes the point that individuals and groups are selected on the basis of their potential to yield new insights into, and enrich, the developing/emergent theory, i.e. the researcher asks whom to turn to next in contributing to the development of the theory.

Theoretical sampling differs from statistical sampling in that: (a) the former does not know in advance what will be the relevant population, whereas the latter does; (b) the former may involve ongoing, new, multiple samples whereas the latter typically does not; (c) the former does not define in advance the sample size, whereas the latter does; (d) in the former the sampling ends when theoretical saturation has been reached whereas in the latter the sampling ends when the whole, predefined sample has been studied; (e) sampling is based on the relevance to the case whereas the latter is based on representativeness (Flick, 2009, pp. 119–21).

Non-probability sampling can be of *people* and of *issues*. Samples of people might be selected because the researcher is concerned to address specific issues, for example, students who misbehave, those who are reluctant to go to school, those with a history of drug dealing, those who prefer extra-curricular to curricular activities. Here it is the issue that drives the sampling, and so the questions become not only 'whom should I sample?' but 'what should I sample?' (Mason, 2002, pp. 127–32). It is not only people who may be sampled, but texts, documents, records, settings, environments, events, objects, organizations, occurrences, activities, and so on.

12.10 Sampling in qualitative research

In qualitative research, often non-probability, purposive samples are employed. However, whilst much of the discussion of probability samples is more relevant to quantitative research (though not exclusively so), and whilst much of the discussion of non-probability samples is more relevant to qualitative research (though not exclusively so), some qualitative research also raises a fundamental question about sampling. The question is this: if sampling presupposes an identifiable population from which a sample is drawn, then is it actually realistic or relevant to identify a population or its sample?

In much qualitative research the emphasis is placed on the uniqueness, the idiographic and exclusive distinctiveness of the phenomenon, group or individuals in question, i.e. they only represent themselves, and nothing or nobody else. In such cases it is perhaps unwise to talk about a 'sample', and more fitting to talk about a group, or individuals. How far they are representative of a wider population or group is irrelevant, as much qualitative research seeks to explore the particular group under study, not to generalize. If, in the process, other groups find that issues raised apply to

them then this is a bonus rather than a necessity, for example, as in case study research.

Further, a corollary of the sympathy between qualitative research and non-probability sampling is that there are no clear rules on the size of the sample in qualitative research; size is informed by 'fitness for purpose', and sample size, therefore, might vary from one to many (Marshall and Rossman, 2016, p. 108). For example, a case study might involve only one child (e.g. Axline, 1964); a grounded theory might continue to add samples until theoretical saturation is reached (i.e. where new data no longer add to the theory construction or themes, or their elements); an ethnography takes in the whole of the group under study, sometimes without any intention of representing a wider population (e.g. Patrick, 1973) and at other times seeking to represent some key features of a wider population (e.g. Willis, 1977). Indeed Flick (2009, p. 123) notes that the basis of choosing sample strategies in qualitative research (including all the non-probability sampling strategies introduced above) is to provide 'rich and relevant information'.

This is not to say that there are no occasions on which, in qualitative research, a sample cannot fairly represent a population. Indeed Onwuegbuzie and Leech (2007, p. 240) argue that external generalizability and inferences to a whole population can feature in qualitative research, and that, as in quantitative research, this typically requires a large sample to be drawn (p. 242). The authors contrast this with internal generalizability, in which data from a sub-group of a sample seeks to be generalizable to the whole sample. That said, they note (p. 249) that, many times, the purpose of the sampling is not to make generalizations, not to make comparisons, but to present unique cases that have their own, intrinsic value.

Onwuegbuzie and Leech (2007, p. 242) suggest that, in qualitative research, the sample size should be large enough to generate 'thick descriptions' (Geertz, 1973) and rich data, though not so large as to prevent this from happening due to data overload or moves towards generalizability, and not so small as to prevent theoretical saturation (discussed earlier) from being achieved. They also counsel (p. 245) that sub-groups in a sample should not be so small as to prevent data redundancy or data saturation, and, in this respect, they recommend that each sub-group should contain no fewer than three cases. As with quantitative data, they note that, as the number of strata increases, so will the size of the sample.

12.11 Sampling in mixed methods research

We introduced sampling in mixed methods research in Chapter 2. We take the discussion further here. Teddlie and Tashakkori (2009, pp. 180–1), drawing on the work of Teddlie and Yu (2007), indicate that it is commonplace for mixed methods research to use more than one kind of sample (probability, non-probability) and to use samples of different sizes, scope and types (cases: people; materials: written, oral observational; other elements in social situations: locations, times, events etc.) within the same piece of research. This harks back to the work of Spradley (1980) on participant observation, Patton (1990) on qualitative research and Miles and Huberman (1994) in discussing actors (participants), settings, events and processes. Even though mixed methods may be used, this does not rule out the fact that, in some mixed methods research, a numerical approach may predominate – with the sampling implications indicated earlier in this chapter (e.g. probability sampling and sample size calculation) – whilst in other mixed methods approaches qualitative data may predominate, with an emphasis on purposive and non-probability sampling (cf. Teddlie and Yu, 2007, p. 85).

Teddlie and Tashakkori (2009, pp. 185–91) provide a useful overview of different mixed methods sampling designs (see also Chapter 2 this volume). In *parallel mixed methods sampling* both probability and non-probability samples are selected, running side by side simultaneously, but separate from each other, i.e. data from one sample do not influence the collection of data from the other and vice versa. Onwuegbuzie and Leech (2007, p. 239) add that parallel sampling designs enable comparisons to be made across two or more sub-groups of a sample that are within the same level of the sample (e.g. girls and boys).

In *sequential mixed methods sampling* (Teddlie and Tashakkori, 2009, pp. 185–91) one kind of sample (both probability and non-probability) precedes another and influences the proceeding sample; in other words, what one gathers from an early sample influences what one does in the next stage with a different sample. For example, numerical data might set the scene for in-depth interviewing, perhaps identifying extreme or deviant cases, critical cases, variables on which the results are either homogeneous or highly varied; alternatively, qualitative data (e.g. case studies or focus groups) might identify issues for exploration in a numerical survey.

In *multilevel mixed methods sampling*, different kinds of sample (both probability and non-probability and either separately or together) are used at different levels of units of analysis, for example: individual

students, classes, schools, local authorities, regions. Onwuegbuzie and Leech (2007, p. 240) suggest that multilevel sampling designs enable comparisons to be made between two or more sub-groups that are drawn from different levels of the study (e.g. individual students and teachers, or individual students and schools, as there is a perceptible hierarchy operating here). They add that this is facilitated by software (e.g. NVivo) that enables such comparative data to be collected and presented by sub-group. They also caution researchers to note that, often, a sub-sample from one level is not the same size as the sub-sample from another (p. 249). For instance, there may be thirty individual students but only one or two teachers for that group of students (they note, in this context, that it is frequently the case that the levels are related, e.g. students and teachers from the same school, rather than being separate, e.g. students from one school and teachers from another).

Teddlie and Tashakkori (2009, p. 191) provide a worked example of a multilevel, mixed methods sampling design in a school effectiveness study, in which:

- at level one, students were selected by probability (random) and purposive sampling (typical cases and complete collection sampling);
- at level two, teachers and classrooms were selected by probability (random and random stratified) and purposive sampling (intensity and typical case sampling);
- at level three, schools were sampled using purposive samples (extreme and deviant case sampling, intensity sampling and typical case sampling);
- at level four, school districts were sampled using probability sampling (cluster samples) and stratified purposive samples;
- at level five, state school systems were sampled using purposive or convenience sampling.

Teddlie and Tashakkori (2009, p. 186) suggest that, in *stratified purposive sampling* the researcher identifies the different strata (e.g. sub-groups) within the population under study, and then selects a limited number of cases from within each of those sub-groups, ensuring that the selection of these cases is based on purposive sampling strategies (i.e. fitness for purpose), drawing on the range of purposive sampling strategies outlined earlier in this chapter. This, they aver, enables the researcher to make comparisons across groups (strata) as required. In this case the purposive sample is a subset of the probability sample (Teddlie and Yu, 2007, p. 93).

Teddlie and Tashakkori (2009, pp. 186–7) also commend *purposeful random sampling*, in which the researcher takes a random sample from a small number of cases from the population (a probability sample) that has already been drawn from a purposive sample (where the population has been chosen for a specific purpose).

Onwuegbuzie and Leech (2007, p. 239) introduce *nested sampling designs*, which enable comparisons to be made between two or more members of the same sub-group and the whole sample. The members of a sub-group represent a sub-sample of the whole sample (p. 246). They give the example (p. 240) of a comparison between key informants and the whole sample.

Teddlie and Tashakkori (2009) also provide useful guidance for sampling in mixed methods research (pp. 192–3), suggesting that the sampling strategy should:

- derive logically from the research questions or hypotheses being investigated/tested;
- be faithful to the assumptions on which the sampling strategies are based (e.g. random allocation, even distributions of characteristics in the population etc.);
- generate qualitative and quantitative data for answering the research questions;
- enable clear inferences to be drawn from both the numerical and qualitative data;
- abide by ethical principles;
- be practicable (able to be done) and efficient;
- enable generalizability of the results (and should indicate to whom the results are generalizable);
- be reported in a level of detail that will enable other researchers to understand it and perhaps use it in the future.

12.12 Planning a sampling strategy

There are several stages in planning the sampling strategy:

Stage 1: Decide whether you need a sample, or whether it is possible to have the whole population.
Stage 2: Identify the population, its important features (the sampling frame) and its size.
Stage 3: Identify the kind of sampling strategy you require (e.g. which variant of probability, non-probability or mixed methods sample you require).
Stage 4: Ensure that access to the sample is guaranteed. If not, be prepared to modify the sampling strategy (stage 3).
Stage 5: For probability sampling, identify the confidence level and confidence intervals that you require. For non-probability sampling, identify the people whom you require in the sample.

Stage 6: Calculate the numbers required in the sample, allowing for non-response, incomplete or spoiled responses, attrition and sample mortality, i.e. build in redundancy by oversampling.
Stage 7: Decide how to gain and manage access and contact (e.g. advertisement, letter, telephone, email, personal visit, personal contacts/friends).
Stage 8: Be prepared to weight (adjust) the data, once collected.

12.13 Conclusion

The message from this chapter is the same as for many of the others, namely, every element of the research should not be arbitrary but planned and deliberate, and the criterion of planning must be 'fitness for purpose'. The selection of a sampling strategy must be governed by the criterion of suitability. The choice of which strategy to follow must be mindful of the purposes of the research, the timescales and constraints on the research, the research design, the methods of data collection and the methodology of the research. The sampling chosen must be appropriate for all of these factors if validity is to be served.

To the question 'how large should my sample be?', the answer is complicated. This chapter has suggested that it all depends on:

- the research purposes, questions and design;
- the size and nature of the population from which the sample is drawn;
- the heterogeneity of the population from which the sample is drawn;
- the confidence level and confidence interval required;
- the likely response rate;
- the accuracy required (the smallest sampling error sought);
- the kinds of variables to be used (categorical, continuous);
- the statistical power required;
- the statistics to be used;
- the scales being used;
- the number of strata required;
- the number of variables included in the study;
- the variability of the factor under study;
- the kind(s) of sample (different kinds of sample within probability, non-probability and mixed methods sampling);
- the representativeness of the population in the sample;
- the allowances to be made for attrition and non-response;
- the need to keep proportionality in a proportionate sample;
- the kind of research that is being undertaken (qualitative/quantitative/mixed methods).

That said, this chapter has urged researchers to use large rather than small samples in quantitative research and sufficiently large and small samples to enable thick descriptions to be achieved in qualitative research. Table 12.4 presents a summary of the types of samples introduced in this chapter.

Decisions on sampling must be made with reference to the criterion of fitness for purpose of the research (internally on the purposes of the study and externally on the intention to generalize or not to generalize), fitness with the research question(s) and match with the focus of the research. Which and how many individuals, groups, communities, institutions, events, places, sites, actions, processes, behaviours etc. to include, and whether to use random sampling (which may or may not provide depth of description and explanation), are complex issues (Marshall and Rossman, 2016, p. 110). How systematic and predetermined or open are the samples depends on the nature of the study. Sampling strategies, as Flick (2009) remarks, describe ways of disclosing and understanding the field (p. 125), and this may require a large, small, wide or narrow sample. Sampling decisions may determine the nature, reliability, validity, credibility, trustworthiness, utility and generalizability of the data collected and, indeed, how to collect such data.

TABLE 12.4 TYPES OF SAMPLE

Probability samples	*Non-probability samples*	*Mixed methods sampling designs*
Simple random sampling	Convenience sampling	Parallel mixed methods sampling
Systematic sampling	Quota sampling	Sequential mixed methods sampling
Random stratified sampling	Purposive sampling:	Multilevel mixed methods sampling
Cluster sampling	Boosted sample	Stratified purposive sampling
Stage sampling	Negative case sampling	Purposeful random sampling
Multi-phase sampling	Typical case sampling	Nested sampling designs
	Extreme/deviant case sampling	
	Intensity sampling	
	Maximum variation sampling	
	Homogeneous sampling	
	Reputational case sampling	
	Revelatory case sampling	
	Critical case sampling	
	Politically important case sampling	
	Complete collection sampling	
	Theoretical sampling	
	Confirming and disconfirming case sampling	
	Opportunistic sampling	
	Snowball sampling	
	Dimensional sampling	
	Volunteer sampling	

Companion Website

The companion website to the book provides PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Sensitive educational research

CHAPTER 13

This chapter addresses several aspects of sensitive research:

- defining sensitive research
- issues of sampling and access
- ethical issues
- effects on the researcher
- researching powerful people
- researching powerless and vulnerable people
- asking questions

It argues that researchers have to be acutely aware of the sensitivities at work in any piece of research that they are undertaking.

13.1 Introduction

All educational research is sensitive or has the potential to become sensitive (cf. Fahie, 2014); the question is one of degree. The researcher has to be sensitive to the context, the cultures, the participants, the consequences of the research on a range of parties (including not only those being researched but also, e.g., researchers, transcribers and readers), the powerless, the powerful, people's agendas and suchlike. Being sensitive is as much about ethics and behaving ethically as it is about the research itself. Researchers have to be very careful on a variety of delicate issues.

The chapter sets out different ways in which educational research might be sensitive. It then takes two significant issues in the planning and conduct of sensitive research – sampling and access – and indicates why these might be challenging for researchers and how they might be addressed. This includes a discussion of gatekeepers and their roles. Sensitive research raises a range of difficult, sometimes intractable, ethical issues; it can also affect researchers and other participants in the research, and we address these here. Investigations involving powerful and powerless people are taken as an instance of sensitive educational research, and this is used to examine several key problematic matters in such research. The chapter moves to a practical note, proffering advice on how to ask questions in sensitive research. Finally, the chapter sets out a range of key issues to be addressed in the planning, conduct and reporting of sensitive research.

13.2 What is sensitive research?

Sensitive research is that 'which potentially poses a substantial threat to those who are involved or have been involved in it' (Lee, 1993, p. 4), when those studied view the research as somehow undesirable (Van Meter, 2000), or when the research generates risk or potential harm for the participants (widely defined) (Corbin and Morse, 2003; Dickson-Swift *et al*., 2007, 2008, 2009; Fahie, 2014; Emerald and Carpenter, 2015). However, sensitivity can derive from many sources, including:

- consequences for the participants (Sieber and Stanley, 1988, p. 49; McCosker *et al*., 2001; Kavanagh *et al*., 2006, p. 245);
- consequences for others, for example, family members, associates, social groups and the wider community, research groups and institutions (Lee, 1993, p. 5), researchers, transcribers and readers (Dickson-Swift *et al*., 2007, 2008, 2009; Fahie, 2014);
- contents, for example, taboo or emotionally charged areas of study (Farberow, 1963), such as criminality, deviance, sex and sexual abuse, race, bereavement, violence, politics, policing, human rights, drugs, poverty, illness, mental health, religion and the sacred, lifestyle, family, finance, physical appearance, power and vested interests (Lee, 1993; Arditti, 2002; Chambers, 2003; Dickson-Swift *et al*., 2007, 2008, 2009; Fahie, 2014);
- situational and contextual circumstances (Lee, 1993);
- intrusion into private, intimate spheres and deep personal experience (Lee and Renzetti, 1993, p. 5), for example, sexual behaviour, religious practices, death and bereavement, even income and age;
- potential sanction, risk or threat of stigmatization, incrimination, costs or career loss to the researcher,

participants or others, for example, groups and communities (Lee and Renzetti, 1993; Renzetti and Lee, 1993; De Laine, 2000), a particular issue for the researcher who studies human sexuality and who, consequently, suffers from 'stigma contagion', i.e. sharing the same stigma as those being studied (Lee, 1993, p. 9);

- impingement on political alignments (Lee, 1993);
- penetration of personal defences, be they of the researched or the researcher (Dickson-Swift *et al*., 2006, 2007, 2008, 2009; Fahie, 2014);
- cultural and cross-cultural factors and inhibitions (Sieber, 1992, p. 129; Tillman, 2002);
- fear of scrutiny and exposure (Payne *et al*., 1980);
- threat to the researcher and to the family members and associates of those studied (Lee, 1993); Lee suggests that 'chilling' may take place, i.e. where researchers are 'deterred from producing or disseminating research' because they anticipate hostile reactions from colleagues, for example, on race or ethnicity (p. 34). 'Guilty knowledge' may bring personal and professional risk from colleagues (De Laine, 2000, p. 67; see also Dickson-Swift *et al*., 2008); it is threatening both to researchers and participants (ibid., p. 84);
- methodologies and conduct, for example, when junior researchers conduct research on powerful people, when men interview women, when senior politicians are involved, and where access and disclosure are difficult (Simons, 1989; Ball, 1990, 1994a; Liebling and Shah, 2001; Walford, 2012).

Sometimes all, or nearly all, of the issues listed above are present simultaneously. Indeed what starts as seemingly innocuous research can turn out to be sensitive (McCosker *et al*., 2001).

In some situations the very activity of actually undertaking educational research per se may be sensitive. This has long been the situation in totalitarian regimes, where permission has typically had to be granted by senior government officers and departments in order to undertake educational research. Closed societies may only permit educational research on approved, typically non-sensitive and comparatively apolitical topics. As Lee (1993, p. 6) suggests: 'research for some groups … is quite literally an anathema'. The very *act* of doing the educational research, regardless of its purpose, focus, methodology or outcome, is itself a sensitive matter (Morrison, 2006). In this situation the conduct of educational research may hinge on interpersonal relations, local politics and micro-politics. What start as being simply methodological issues can turn out to be ethical and political/micro-political minefields.

Lee (1993, p. 4) suggests that sensitive research falls into three main areas: (a) intrusive threat (probing into areas which are 'private, stressful or sacred'); (b) studies of deviance and social control, i.e. which could reveal information that could stigmatize or incriminate (threat of sanction); and (c) political alignments, revealing the vested interests of 'powerful persons or institutions, or the exercise of coercion or domination', or extremes of wealth and status (Lee, 1993). As Beynon (1988, p. 23) says, 'the rich and powerful have encouraged hagiography, not critical investigation'.

Lee (1993, p. 8) argues that there has been a tendency to 'study down' rather than 'study up', i.e. to direct attention to powerless rather than powerful groups, not least because these are sometimes easier and less sensitive to investigate. Sensitive educational research can act as a voice for the weak, the oppressed, those without a voice or who are not listened to; equally it can focus on the powerful and those in high-profile positions.

The three kinds of sensitivities indicated above, (a), (b) and (c), may appear separately or in combination. The sensitivity concerns not only the topic itself, but, often more importantly, 'the relationship between that topic and the social context' within which the research is conducted (Lee, 1993, p. 5). What appears innocent to the researcher may be highly sensitive to the researched or to other parties. Threat is a major source of sensitivity; indeed Lee (p. 5) suggests that, rather than generating a list of sensitive topics, it is more fruitful to look at the conditions under which 'sensitivity' arises within the research process. Given this issue, the researcher will need to consider how sensitive the educational research will be, not only in terms of the subject matter itself, but also in terms of the several parties that have a stake in it, for example: headteachers/principals and senior staff; parents; students; schools; governors; local politicians and policy makers; the researcher(s) and research community; government officers; the community; social workers and school counsellors; sponsors and members of the public; members of the community being studied; and so on.

Sensitivity inheres both in the educational topic under study, but also, much more significantly, in the social context in which the educational research takes place and on the likely consequences of that research on all parties. Doing research is not only a matter of designing a project and collecting, analysing and reporting data – that is the optimism of idealism or ignorance; it is also a matter of interpersonal relations, potentially continual negotiation, delicate forging and sustaining of relationships, setbacks, modification and

compromise. In an ideal world educational researchers would be able to plan and conduct their studies untrammelled; however, this typically does not happen in the real world, and sensitive educational research exposes this very clearly. Whilst most educational research will incur sensitivities, the benefit of discussing sensitive research per se is that it highlights what these delicate issues might be and how they might be felt at their sharpest. We advise readers to consider most educational research as sensitive, to anticipate what those sensitivities might be and what trade-offs might be necessary.

13.3 Sampling and access

Lee (1993, p. 60) suggests that there are potentially serious difficulties in sampling and access in sensitive research, not least because of the problem of estimating the size of the population from which the sample is to be drawn, as members of particular groups, for example, deviant or clandestine groups, will not want to disclose their associations. Similarly, like-minded groups may not wish to open themselves to public scrutiny. They may have much to lose by revealing their membership and, indeed, their activities may be illicit, critical of others, unpopular, threatening to their own professional security, deviant and less frequent than activities in other groups, making access a major obstacle. What if a researcher is researching truancy, or teenage pregnancy, or bullying, or solvent abuse among school students, or alcohol and medication use among teachers, or family relationship problems brought about by stress in teaching?

Lee (1993) suggests several strategies to be used (p. 61), either separately or in combination, for sampling 'special' populations (e.g. rare or deviant populations):

- *List sampling*: looking through public domain lists of, for example, the recently divorced (though such lists may be more helpful to social researchers than, specifically, educational researchers).
- *Multi-purposing*: using an existing survey to reach populations of interest (though problems of confidentiality may prevent this from being employed).
- *Screening*: targeting a particular location and canvassing within it (which may require much effort for little return).
- *Outcropping*: going to a particular location where known members of the target group congregate or can be found (e.g. Humphreys' celebrated study of homosexual 'tearoom trade' in 1970); in education this may be a particular staffroom (for teachers), or meeting place for students. Outcropping risks bias, as there is no simple check for representativeness of the sample.
- *Servicing*: Lee (1993, p. 72) suggests that it may be possible to reach research participants by offering them some sort of benefit or service in return for their participation. Researchers must be certain that they really are able to provide the services promised.
- *Professional informants*: Lee (1993, p. 73) suggests these could be, for example, police, doctors, priests or other professionals. In education these may include social workers and counsellors. This may be unrealistic optimism, as these very people may be bound by terms of legal or ethical confidentiality or voluntary self-censorship (e.g. an AIDS counsellor, after a harrowing day at work, may not wish to continue talking to a stranger about AIDS counselling, or a social worker or counsellor may be constrained by professional confidentiality, or an exhausted teacher may not wish to talk about her teaching difficulties). Further, even if such people agree to participate, they may not know the full story (cf. Walford, 2012). Lee gives the example of drug users (p. 73), whose contacts with the police may be very different from their contacts with doctors or social workers, or, the corollary of this, the police, doctors and social workers may not see the same group of drug users.
- *Advertising*: though this can potentially reach a wide population, it may be difficult to control the nature of those who respond, in terms of representativeness or suitability (a particular issue in online research, e.g. surveys).
- *Networking*: this is akin to snowball sampling (see Chapter 12), where one set of contacts puts the researcher in touch with more contacts, who, in turn, put the researcher in touch with yet more contacts and so on. This is a widely used technique, though Lee (1993, p. 66) reports that it is not always easy for contacts to be passed on, as initial informants may be unwilling to divulge members of a close-knit community. On the other hand, Morrison (2006) reports that networking is a popular technique where it is difficult to penetrate a formal organization such as a school, if the gatekeepers (those who can grant or prevent access to others, e.g. the headteacher or senior staff) refuse access. He reports the extensive use of informal networks by researchers, in order to contact friends and professional associates, and, in turn, their friends and professional associates, thereby sidestepping the formal lines of contact through schools.

Hammersley and Atkinson (1983, p. 54) suggest that gaining access is a practical matter and it provides insights into the 'social organisation of the setting'. Walford (2001, p. 33, 2012) argues that gaining access and becoming accepted is a slow process. He sets out a four-stage process of gaining access (2001, pp. 36–47):

Stage 1: *Approach* (gaining entry, perhaps through a mutual friend or colleague – a link person). Walford cautions that an initial letter should only be used to gain an initial interview or an appointment, or even to arrange to telephone the headteacher in order to arrange an interview, not to conduct the research or to gain access.
Stage 2: *Interest* (using a telephone call to arrange an initial interview). Here Walford notes (p. 43) that headteachers may like to talk, and so it is important to let them talk, even on the telephone when arranging an interview to discuss the research.
Stage 3: *Desire* (overcoming objections and stressing the benefits of the research). As Walford wisely comments (p. 44): 'after all, schools have purposes other than to act as research sites'. He makes the telling point that the research may actually benefit the school, but that the school may not realize this until it is pointed out. For example, a headteacher may wish to confide in a researcher; teachers may benefit from discussions with a researcher; students may benefit from being asked about their learning.
Stage 4: *Sale* (where the participants agree to the research).

Whitty and Edwards (1994, p. 22) argue that in order to overcome problems of access, ingenuity and even subterfuge could be considered: 'denied co-operation initially by an independent school, we occasionally contacted some parents through their child's primary school and then told the independent schools we already were getting some information about their pupils'. They also add that it is sometimes necessary for researchers to indicate that they are 'on the same side' as those being researched.[1] Indeed they report that 'we were questioned often about our own views, and there were times when to be viewed suspiciously from one side proved helpful in gaining access to the other' (p. 22). This harks back to Becker's (1967) advice to researchers to decide whose side they are on (cf. Hammersley, 2000).

The use of snowball sampling builds in 'security' (Lee, 1993), as the contacts are those who are known and trusted by the members of the 'snowball'. That said, this itself can lead to bias, as relationships between participants in the sample may consist of 'reciprocity and transitivity' (p. 67), i.e. participants may have close relationships with one another and may not wish to break these. Thus homogeneity of the sample's attributes may result.

Snowball sampling may alter the research, for example changing random, stratified or proportionate sampling into convenience sampling, thereby compromising generalizability or generating the need to gain generalizability by synthesizing many case studies. Nevertheless, it often comes to a choice between accepting non-probability strategies or doing nothing.

Issues of access to people in order to conduct sensitive research may require researchers to demonstrate a great deal of ingenuity and forethought in their planning. Investigators have to be adroit in anticipating problems of access, and set up their studies in ways that circumvent such problems and prevent them from arising in the first place, for example, by exploring their own institutions or personal situations, even if this compromises generalizability. Such anticipatory behaviour can lead to a glut of case studies, action research and accounts of their own institutions, as these are the only kinds of research possible, given the problem of access.

Gatekeepers

Access might be gained through gatekeepers, that is, those who control access. Lee (1993, p. 123) suggests that 'social access crucially depends on establishing *interpersonal trust*'. Gatekeepers play a significant role in research, particularly in ethnographic research (Miller and Bell, 2002, p. 53), as they control access and re-access (p. 55). They may provide or block access; they may steer the course of a piece of research, 'shepherding the fieldworker in one direction or another' (Hammersley and Atkinson, 1983, p. 65), or exercise surveillance over the research.

Gatekeepers may wish to avoid, contain, spread or control risk and therefore may bar access or make access conditional. Making research conditional may require researchers to change the nature of their original plans in terms of methodology, sampling, focus, dissemination, reliability and validity, reporting and control of data (Morrison, 2006). Morrison (2006) found that in conducting sensitive educational research, there were problems of:

- gaining access to schools and teachers;
- gaining permission to conduct the research (e.g. from school principals): resentment by principals;
- people vetting which data could be used;
- finding enough willing participants for the sample;

- schools/institutions/people not wishing to divulge information about themselves;
- schools/institutions not wishing to be identifiable, even with protections guaranteed;
- local political factors that impinged on the school/educational institution;
- teachers'/participants' fear of being identified/traceable, even with protections guaranteed (e.g. if they raised critical matters about the school or others they could lose their contracts);
- unwillingness of teachers to be involved because of their workload;
- the principal deciding on whether to involve the staff, without consulting the staff;
- schools' fear of criticism/loss of face or reputation;
- the sensitivity of the research – the issues being investigated;
- the power/position of the researcher (e.g. if the researcher is a junior or senior member of staff or an influential person in education).

Risk reduction may result in participants imposing conditions on research (e.g. on what information investigators may or may not use; to whom the data can be shown; what is 'public'; what is 'off the record' (and what should be done with off-the-record remarks)). It may also lead to surveillance/'chaperoning' of the researcher whilst the study is being conducted on site (Lee, 1993, p. 125).

Gatekeepers may want to 'inspect, modify or suppress the published products of the research' (Lee, 1993, p. 128). They may also wish to use the research for their own ends, i.e. their involvement may not be selfless or disinterested, or they may want something in return, for example, for the researcher to include in the study an area of interest to the gatekeeper, or to report directly – and maybe exclusively – to the gatekeeper. The researcher has to negotiate a potential minefield here, for example, not to be seen as an informer for the headteacher. As Walford (2001, p. 45) writes: 'headteachers [may] suggest that researchers observe certain teachers whom they want information about'. Researchers may need to reassure participants that their data will not be given to the headteacher.

On the other hand, Lee (1993, p. 127) suggests that the researcher may have to make a few concessions in order to be able to undertake the investigation, i.e. that it is better to do a little of the gatekeeper's bidding rather than not to be able to do the research at all (cf. Morrison, 2006).

In addition to gatekeepers, the researcher may find a 'sponsor' in the group being studied. A sponsor may provide access, information and support. A celebrated example of this is in the figure of 'Doc' in Whyte's classic study of *Street Corner Society* (1993; original study published 1943). Here Doc, a leading gang figure in the Chicago street corner society, is quoted as saying:

> You tell me what you want me to see, and we'll arrange it. When you want some information, I'll ask for it, and you listen. When you want to find out their philosophy of life, I'll start an argument and get it for you.… You won't have any trouble. You come in as a friend.
>
> (Whyte, 1993, p. 292)

As Whyte writes:

> My relationship with Doc changed rapidly.… At first he was simply a key informant – and also my sponsor. As we spent more time together, I ceased to treat him as a passive informant. I discussed with him quite frankly what I was trying to do, what problems were puzzling me, and so on … so that Doc became, in a real sense, a collaborator in the research.
>
> (Whyte, 1993, p. 301)

Whyte comments on how Doc was able to give him advice on how best to behave when meeting people as part of the research:

> Go easy on that 'who', 'what', 'why', 'when', 'where' stuff, Bill. You ask those questions and people will clam up on you. If people accept you, you can just hang around, and you'll learn the answers in the long run without even having to ask the questions.
>
> (Whyte, 1993, p. 303)

Indeed Doc played a role in the writing of the research: 'As I wrote, I showed the various parts to Doc and went over them in detail. His criticisms were invaluable in my revision' (p. 341). In his 1993 edition, Whyte reflects on the study with the question as to whether he exploited Doc (p. 362); it is a salutary reminder of the essential reciprocity that might be involved in conducting sensitive research.

In addressing issues of sampling and access, there are several points that arise from the discussion (Box 13.1).

Much research stands or falls on the sampling. Rather than barring the research altogether, compromises may have to be reached in sampling and access. It may be better to compromise rather than to abandon the research altogether.

BOX 13.1 ISSUES OF SAMPLING AND ACCESS IN SENSITIVE RESEARCH

- How to calculate the population and sample.
- How representative of the population the sample may or may not be.
- What kind of sample is desirable (e.g. random), but what kind may be the only sort that is practicable (e.g. snowball).
- How to use networks for reaching the sample, and what kinds of networks to utilize.
- How to research in a situation of threat to the participants (including the researcher).
- How to protect identities and threatened groups.
- How to contact the hard-to-reach.
- How to secure and sustain access.
- How to find and involve gatekeepers and sponsors.
- What to offer gatekeepers and sponsors.
- On what matters compromise may need to be negotiated.
- On what matters can there be no compromise.
- How to negotiate entry and sustained field relations.
- What services the researcher may provide.
- How to manage initial contacts with potential groups for study.

13.4 Ethical issues in sensitive research

A difficulty arises in sensitive research in that the researcher can be party to 'guilty knowledge' (De Laine, 2000) and have 'dirty hands' (Klockars, 1979) about deviant groups or members of a school who may be harbouring counter-attitudes to those prevailing in the school's declared mission. Pushed further, this means that the researcher will need to decide the limits of tolerance, beyond which he/she will not venture. For example, in Patrick's (1973) study of a Glasgow gang, the researcher is witness to a murder. Should he report the matter to the police and, thereby, 'blow his cover', or remain silent in order to keep contact with the gang, thereby breaking the law which requires a murder to be reported?

In interviewing students, they may reveal sensitive matters about themselves, their family or their teachers, and the researcher will need to decide whether and how to act on this kind of information. What should the researcher do, for example, if, during the course of an interview with a teacher about the leadership of the headteacher, the interviewee indicates that the headteacher has had sexual relations with a parent, or has an alcohol problem? Does the researcher, in such cases, do nothing in order to gain research knowledge, or does he act? What is in the public interest – the protection of an individual participant's private life, or the interests of the researcher? Indeed Lee (1993, p. 139) suggests that some participants may even deliberately engineer situations whereby the researcher gains 'guilty knowledge' in order to test the researcher's affinities: 'trust tests'.

Ethical issues are thrown into sharp relief in sensitive educational research. The question of covert research rises to the fore, as the study of deviant or sensitive situations may require the researcher to go under cover in order to obtain data. Access is often a serious problem in educational and social research (Munro *et al*., 2004, p. 295), particularly if such access is controlled by powerful people (Morrison, 2006). Powerful gatekeepers may control several aspects of participants' lives (Munro *et al*., 2004, p. 302) such as promotion, in-service training and work allocations, and it may be necessary to consider covert research or deception. Covert research may overcome 'problems of reactivity' (Lee, 1993, p. 143), wherein the research influences the behaviour of the participants (Hammersley and Atkinson, 1983, p. 71). Deception, though questioned in codes of practice for educational research (see Chapter 7), is not ruled out in these same codes, and there may be cases where the violation of informed consent, or telling lies, or not disclosing that one is conducting research, may be considered to be justified in order to obtain data on honest, natural behaviours, views or practices. If a researcher seeks the informed consent of violent teachers to study their violent behaviour, is there any real likelihood that the research will actually take place, whereas if one asks permission to study the behaviour of the students in their class, and keeps quiet about the real purpose which is to study violent teachers, is it more likely that access will be granted? And yet, surely, it is important in the interests of the

students, the school, even the violent teacher themselves, that the problem be exposed and be evidence-based?

Covert research or deliberate deception may also enable the researcher to obtain insiders' true views, for, without the cover of those being researched not knowing that they are being studied, entry could easily be denied, and access to important areas of understanding could be lost. This is particularly so in the case of researching powerful people who may not wish to disclose information and who, therefore, may prevent or deny access. The ethical issue of informed consent, in this case, is violated in the interests of exposing matters that are in the public interest.

To the charge that this is akin to spying, Mitchell (1993, p. 46) makes it clear that there is a vast difference between covert research and spying:

- Spies, he argues, seek to further a particular value system or ideology; research seeks to understand rather than to persuade.
- Spies have a sense of mission and try to achieve certain instrumental ends, whereas research has no such specific mission.
- Spies believe that they are morally superior to their subjects, whereas researchers have no such feelings; indeed, with reflexivity being so important, they are sensitive to how their own role in the investigation may distort the research.
- Spies are supported by institutions which train them to behave in certain ways of subterfuge, whereas researchers have no such training.
- Spies are paid to do the work, whereas researchers often operate on a not-for-profit or individualistic basis.

On the other hand, not to gain informed consent could lead to participants feeling duped, very angry, used and exploited when the results of the research are eventually published and they realize that they have been studied without their approval or informed consent.[2] The researcher is seen as a predator (Lee, 1993, p. 157), using the research 'as a vehicle for status, income or professional advancement which is denied to those studied'. As Lee remarks (p. 157), 'it is not unknown for residents in some ghetto areas of the United States to complain wryly that they have put dozens of students through graduate school'. Further, the researched may: have no easy right of reply; feel misrepresented by the research; feel that they have been denied a voice; have wished not to be identified and their situation put into the public arena; feel that they have been exploited.

The cloak of anonymity is often vital in sensitive research, such that respondents are entirely untraceable. This raises the issue of 'deductive disclosure' (Boruch and Cecil, 1979), wherein it is possible to identify the individuals (people, schools, departments, etc.) in question by reconstructing and combining data. Researchers should guard against this possibility. Where the details that are presented could enable identification of a person (e.g. in a study of a school there may be only one male teacher aged fifty who teaches biology, such that putting a name is unnecessary, as he will be identifiable), it may be incumbent on the researcher not to disclose such details, so that readers, even if they wished to reassemble the details in order to identify the respondent, are unable to do so.

The researcher may wish to preserve confidentiality and non-traceability, but may also wish to be able to gather data from individuals on more than one occasion. In this case a 'linked file' system (Lee, 1993, p. 173) can be employed. Here three files are kept; in the first file the data are held and arbitrary numbers are assigned to each participant; the second file contains the list of respondents; the third file contains the list information necessary to be able to link the arbitrarily assigned numbers from the first file to the names of the respondents in the second, and this third file is kept by a neutral 'broker', not the researcher. This procedure is akin to double-blind clinical experiments, in which the researcher does not know the names of those who are or are not receiving experimental medication or a placebo. That this may be easier in respect of quantitative rather than qualitative data is acknowledged by Lee (1993, p. 179).

Clearly, in some cases, it is impossible for individual people, schools and departments not to be identified, for example, schools may be highly distinctive and, therefore, identifiable (Whitty and Edwards, 1994, p. 22). In such cases clearance may need to be obtained for the disclosure of information. This is not as straightforward as it may seem. For example, a general principle of educational research is that no individuals should be harmed (non-maleficence), but what if a matter that is in the legitimate public interest is brought to light (e.g. a school's failure to keep to proper accounting procedures)? Should the researcher follow up the matter privately, publicly or not at all? If it is followed up then certainly harm may come to the school's officers.

Ethical issues in the conduct of research are thrown into sharp relief against a backdrop of personal, institutional and societal politics, and the boundaries between public and private spheres are not only relative but ambiguous. The ethical debate is heightened, for example in the potential tension between the individual's right to privacy versus the public's right to know,

and the concern not to damage or harm individuals versus the need to serve the public good. Because public and private spheres may merge, it is difficult, if not impossible, to resolve such tensions straightforwardly (cf. Day, 1985; Lee, 1993). As Walford (2001, p. 30) writes: 'the potential gain to public interest … was great. There would be some intrusion into the private lives of those involved, but this could be justified in research on … an important policy issue'. The end justified the means.

These issues are felt most sharply if the research risks revealing negative findings. To expose practices to research scrutiny may be like taking the plaster off an open wound (Wood, 1980). What responsibility to the research community does the researcher have? If a negative research report is released, will schools retrench, preventing future research in schools from being undertaken (a particular problem if the researcher wishes to return or wishes not to prevent further researchers from gaining access)? Whom is the researcher serving – the public, the schools, the research community? The sympathies of the researcher may be called into question here; politics and ethics may be uncomfortable bedfellows in such circumstances. Research data, such as the negative hidden curriculum of training for conformity in schools (Morrison, 2009) may not endear researchers to schools.

This can risk stifling educational research – it is simply not worth the personal or public cost. As Simons (2000, p. 45) says: 'the price is too high'.

Further, Mitchell (1993) writes that adhering to privacy may lead to 'timorous social scientists' excusing themselves from risks associated with confronting powerful people, the privileged and self-protecting groups who may not wish to disclose their actions to the scrutiny of the public (p. 54) (see also Lee, 1993, p. 8). Researchers may not wish to risk offending the powerful or placing themselves in uncomfortable situations. As Simons and Usher (2000, p. 5) remark: 'politics and ethics are inextricably entwined'.

In private, students and teachers may criticize their own schools, for example, in terms of management, leadership, work overload and stress, but they may be reluctant to do so in public, and indeed teachers who are on renewable contracts will not bite the hand that feeds them; they may say nothing rather than criticize (Burgess, 1993a; Morrison, 2002b).

The field of ethics in sensitive research may be different from ethics in everyday research, in significance rather than range of focus. The same issues faced in all educational research are addressed here, and we advise readers to review Chapter 7 on ethics. However, sensitive research highlights particular ethical issues very sharply, as presented in Box 13.2.

These are only introductory issues. We refer the reader to Chapter 7 for further discussion of these and other ethical issues. The difficulty with ethical issues is that they are 'situated' (Simons and Usher, 2000), i.e. contingent on specific local circumstances and situations. They have to be negotiated and worked out in

BOX 13.2 ETHICAL ISSUES IN SENSITIVE RESEARCH

- How does the researcher handle 'guilty knowledge' and 'dirty hands'?
- Whose side is the researcher on? Does this need to be disclosed? What if the researcher is not on the side of the researched?
- When are covert research or deception justified?
- When is the lack of informed consent justified?
- Is covert research spying?
- How should the researcher overcome the charge of exploiting the participants (i.e. treating them as objects instead of as subjects of research)?
- How should the researcher address confidentiality and anonymity?
- How should the balance be struck between the individual's right to privacy and the public's right to know?
- What is really in the public interest?
- How to handle the situation where it is unavoidable to identify participants?
- What responsibility does the researcher have to the research community, some of whom may wish to conduct further research in the field?
- How does the researcher handle frightened or threatened groups who may reveal little?
- What protections are in the research, for whom, and from what?
- What obligations does the researcher have?

relation to the specifics of the situation; universal guidelines may help but they don't usually solve the practical problems; they have to be interpreted locally.

13.5 Effects of sensitive research on the researcher

Sensitive research can take its toll on several parties: those who are being researched, researchers, transcribers, supervisors, examiners and, indeed, readers (Dickson-Swift *et al*., 2007, 2008, 2009; McCosker *et al*., 2001; Fahie, 2014). Here the earlier definition from Lee (1993) as that 'which potentially poses a substantial threat to those who are involved or have been involved in it' (p. 4) applies not only to those being researched but to other parties who might be affected by the research. Fahie (2014), for example, reporting a study of workplace bullying in primary schools, notes the potential risk to the researcher here, commenting that one research participant managed to obtain the personal contact details of the researcher and telephoned him some 40–50 times over the course of one year, intruding into his personal life.

Let us say that the researcher is faced by a teenager who sobs uncontrollably when recounting her genuinely dreadful account of childhood abuse, which really touches the researcher, the transcriber of the research interview (Dickson-Swift *et al*., 2007) and indeed the reader? Can they or should they show or not some kind of empathy, indeed can they prevent themselves from having and showing a deep emotional reaction? Emotional and cognitive actions and reactions are not as separable as we might find convenient (e.g. Dickson-Swift *et al*., 2007, 2008, 2009; Fahie, 2014), and indeed research is often an emotional experience.

Researching sensitive topics can be regarded as 'emotion work', i.e. that kind of activity which involves the management of emotions as an important element of work (in this case, of educational research) (Dickson-Swift *et al*., 2009; Hochschild, 2012). This typically includes work which involves much face-to-face or voice-to-voice interaction, particularly with those who are external to the organization as well as those who are internal to it, and which requires workers to produce an emotional state in others whilst managing their own emotions (p. 63).

As emotion workers, researchers have to manage their own emotions, yet emotions are fundamental to being human, and this poses a challenge: should the researcher remain emotionally relatively aloof and distant from the person, say, being interviewed, in order to maintain scientific or researcher objectivity, or should they allow their own emotions to be part of the process of, say, interviewing? Should they hold back or show their emotions? Indeed is it really possible to hold back if one is moved to tears? Researchers may not be able to stop themselves here, but is this acceptable (Dickson-Swift *et al*., 2007, 2008, 2009; Fahie, 2014)?

There are different responses to this: some would argue that it is perfectly acceptable for researchers to show their emotions, not least as, being perhaps coldly instrumental, this might stimulate an even richer response from those being researched. Further, it is important to respond to a research participant in human terms, and if this means not holding back the researcher's tears, anger or sadness, then so be it; as Ely *et al*. (1991) remark, if researchers are to study humans, then they have to be ready to 'face human feelings' (p. 49) and respond to the research participants as human beings, not robots, would respond.

When researching sensitive topics, natural empathy might establish a bond, a connection or rapport, between the researcher and the researched. Such reciprocity recognizes the essential humanity of a human situation (Dickson-Swift *et al.*, 2009). Indeed, when researching the marginalized and vulnerable, the research might be the only opportunity that they have had to tell their story to anyone, and for the researcher to show his/her emotional involvement here might support the catharsis that such participant disclosure might value (Dickson-Swift *et al.*, 2007).

By contrast, others would argue that for the researcher to introduce his or her own emotions onto an already emotionally charged, intense situation is somehow unworthy, improper, unscientific and a threat to rigour, sending inappropriate signals to the person being researched (or indeed even to his or her academic colleagues (Dickson-Swift *et al*., 2009)) and that any emotions should be held in check at least until after the encounter.

At issue here is the recognition that doing sensitive, emotionally charged research exacts its price on researchers (Dickson-Swift *et al*., 2007, 2008, 2009; Fahie, 2014). For example they may:

- feel emotionally and physically exhausted, become emotionally hardened and desensitized, for example, no longer able to be shocked;
- experience insomnia, nightmares, permanent tiredness and depression;
- feel guilty or angry in reporting but not taking action to alleviate or remediate the participant's situation;
- feel guilty in having affected the research participant;
- feel vulnerable (to their own emotions or to learning something about themselves);

- feel a failure or frustrated in not having managed to control their own emotions or not having maintained boundaries between themselves and the participants and becoming too friendly or empathetic;
- feel guilty at having entered intimately into the lives of others and then leaving them, i.e. a breach of trust, using others as a means to an end;
- feel that the establishing of rapport, indeed friendship, was somehow deceitful, for obtaining data only, again using others as a means to an end;
- feel that the research participants may not want to hear the self-disclosures of the researchers, as this could burden them even more;
- feel that they have let themselves down in breaching their own intention of not being too empathetic or emotional in the research situation (e.g. in an interview);
- have blurred the distinction between research and therapy;
- have failed to protect the research participants;
- feel that they, as keepers of secrets and private, privileged information, have betrayed the trust of the research participant. (Dickson-Swift *et al*. (2007) liken the trust and keeping of secrets to a religious confessional, thereby offending their own conscience.)

Dickson-Swift *et al*. (2007) note that, for some researchers, undertaking sensitive research can become a life-changing experience (p. 342) or an intense emotional, even traumatic encounter. Fahie (2014) illustrates this well, commenting on an interviewee recounting her story of being bullied by a school principal:

> Watching her cry in her own sitting room, listening to her describe the ritual humiliation she encountered in her place of work, and seeing her hands shake as she recalled the vitriolic abuse at the hands of her school principal, impacted upon me deeply by drawing me into the narrative. And I felt angry … the sheer injustice of it and unfairness of her experiences disturbed me profoundly, as did my own inability to 'make it better'. This impotence made me feel frustrated and helpless, as if, in some way, I had left Ann down.
>
> (Fahie, 2014, p. 25)

Here it is not enough simply to state that, ethically speaking, the research must not leave the participants worse off than before the research; rather it is to say that, in addressing sensitive research, care has to be given to support the researchers as well, even as a matter of Health and Safety requirements, both physically and psychologically, be this through, for example, counselling and support staff and services, peer support, mentoring and supervision, security services, social support or suchlike (McCosker *et al*., 2001). In this respect, ethics committees should also consider the possible effects of the research on all parties involved, including often-overlooked parties such as researchers, supervisors, transcribers and other members of the contact circle of those being researched.

McCosker *et al.* (2001) and Fahie (2014) also give practical advice for researchers conducting sensitive research, including: non-disclosure of personal details and personal contact details; conducting interviews in public places and informing another party of the likely starting and finishing times; checking the environment before agreeing the location of the interview; using a different SIM card from one's main SIM card in cellphone conversations with research participants; keeping a record of the time, place and duration of the interview; discussing and conducting debriefings on the research with a mentor and/or supervisor; closely monitoring the emotional impact of the research on the participants; consider spacing out the timing of interviews and the subsequent listening to recordings of interviews on sensitive topics, for example, only a limited number per week, in order to enable researchers not to be emotionally overwhelmed by, or desensitized by, emotionally charged interviews.

13.6 Researching powerful people

A branch of sensitive research concerns that which is conducted on, or with, powerful people, those in key positions, or elite institutions. In education, for example, this could include headteachers/principals and senior teachers, politicians, senior civil servants, decision makers, local authority officers and school governors. This is particularly the case in respect of research on policy and leadership issues (Walford, 1994a, p. 3, 2012). Researching the powerful is an example of 'researching up' rather than the more conventional 'researching down' (e.g. researching children, teachers, student teachers).

What makes the research sensitive is that it is often dealing with key issues of policy generation and decision making, or issues about which there are high-profile debate and contestation, or issues of a politically sensitive nature. Policy-related research is sensitive. This can also be one of the reasons why access is frequently refused. The powerful are those who exert control to secure what they want or can achieve, those with great responsibility and whose decisions have

significant effects on large numbers of people. Indeed they have considerable power in blocking access for researchers, thereby stopping the research, particularly if the issue is controversial or sensitive (e.g. contested fiercely by various parties) (Walford, 2012, p. 112).

Academic educational research on the powerful may be unlike other forms of educational research in that confidentiality may not be able to be assured. The participants are identifiable and public figures. This may produce 'problems of censorship and self-censorship' (Walford, 1994c, p. 229). It also means that information given in confidence and 'off the record' unfortunately may have to remain so. One issue raised in researching the powerful is the disclosure of identities, particularly if it is unclear what has been said 'on the record' and 'off the record' (Fitz and Halpin, 1994, pp. 35–6).

Fitz and Halpin (1994) indicate that the government minister whom they interviewed stated, at the start of the interview, what was to be attributable. They also report that they used semi-structured interviews in their research of powerful people, valuing both the structure and the flexibility of this type of interview, and that they gained permission to record the interviews for later transcription, for the sake of a research record. They also used two interviewers for each session, one to conduct the main part of the interview and the other to take notes (p. 47) and ask supplementary questions, helping to negotiate the way through the interview in which advisers to the interviewee were also present to monitor the proceedings and interject where it was deemed fitting (p. 44). Having two interviewers present also enabled a post-interview cross-check to be undertaken.

Fitz and Halpin comment on the considerable amount of gatekeeping that was present in researching the powerful (p. 40), in terms of access to people (with officers guarding entrances and administrators deciding whether interviews will take place), places ('élite settings'), timing (and scarcity of time with busy respondents), 'conventions that screen off the routines of policy-making from the public and the academic gaze' (p. 48), conditional access and conduct of the research ('boundary maintenance'; p. 49), monitoring and availability. Gewirtz and Ozga (1994, pp. 192–3) suggest that gatekeeping in researching the powerful can produce difficulties which include 'misrepresentation of the research intention, loss of researcher control, mediation of the research process, compromise and researcher dependence'.

Research with powerful people usually takes place on their territory, under their conditions and agendas (a 'distinctive civil service voice'; Fitz and Halpin, 1994, p. 42), working within discourses set by the powerful (and, in part, reproduced by the researchers; p. 40), and with protocols concerning what may or may not be disclosed (e.g. under a government's Official Secrets Act or privileged information), within a world which may be unfamiliar and, thereby, disconcerting for researchers and with participants who may be overly assertive, sometimes making the researcher have to pretend to know less than he or she actually knows. As Fitz and Halpin (1994, p. 40) commented: 'we glimpsed an unfamiliar world that was only ever partially revealed', and one in which they did not always feel comfortable. Similarly, Ball (1994b, p. 113) suggests that 'we need to recognize … the interview as an extension of the "play of power" rather than separate from it, merely a commentary upon it', and that, when interviewing powerful people, 'the interview is both an ethnographic … and a political event'. As Walford remarks:

> Those in power are well used to their ideas being taken notice of. They are well able to deal with interviewers, to answer and avoid particular questions to suit their own ends, and to present their own role in events in a favourable light. They are aware of what academic research involves, and are familiar with being interviewed and having their words tape-recorded. In sum, their power in the educational world is echoed in the interview situation, and interviews pose little threat to their own positions.
>
> (Walford, 1994c, p. 225)

McHugh (1994) comments that access to powerful people may take place not only through formal channels but through intermediaries who introduce researchers to them (p. 55). Here his own vocation as a priest helped him to gain access to powerful Christian policy makers and, as he was advised, 'if you say whom you have met, they'll know you are not a way-out person who will distort what they say' (p. 56). Access is a significant concern in researching the powerful, particularly if the issues being researched are controversial or contested (Walford, 2012).

Access may be difficult, because the very person whom the researcher wishes to meet may be busy or constrained by what he or she may or may not disclose, and the whole point of the meeting is to meet that particular person and not a substitute (cf. Walford, 2012, p. 115). Walford (1994c, p. 222) suggests that access can be eased through informal and personal 'behind the scenes' contacts: 'the more sponsorship that can be obtained, the better' (p. 223), be it institutional or personal. As he also remarks: '[o]ne obvious way of easing access is exploiting pre-existing links with those in power' (Walford, 2012, p. 112). Access can also be eased if the research is seen to be 'harmless' (p. 112); here he reports that female researchers may be at an advantage in that they are

viewed as more harmless and non-threatening (p. 112), particularly, he avers, if they are relatively young and not in a senior position in their own institution (though he also notes research which suggests that a female may not be 'taken as seriously as a male researcher'; p. 112). He also notes that gaining access to powerful people who have retired is easier than those who are still in office (p. 112), though the researcher would have to exercise caution here as the person may be seeking to 'write themselves into history' (p. 112). Walford (1994c) also makes the point that 'persistence pays' (p. 224); as he writes elsewhere (Walford, 2012, p. 115), 'access is a process rather than a one-off decision'.

McHugh (1994) reports the need for meticulous preparation for an interview with the powerful person, to understand the full picture and to be as fully informed as the interviewee, in terms of facts, information and terminology, so that it is an exchange between the informed rather than an airing of ignorance, i.e. to do one's homework. He also states the need for the interview questions to be thoroughly planned and prepared, with very careful framing of questions. He suggests (p. 60) that during the interview it is important for the interviewer to be as flexible as possible, to follow the train of thought of the respondent, but also to be persistent (p. 62) if the interviewee does not address the issue. However, he reminds us that 'an interview is of course not a courtroom' (p. 62) and so tact, diplomacy and – importantly – empathy are essential. Diplomacy in great measure is necessary when tackling powerful people about issues that might reveal their failure or incompetence, and powerful people may wish to control which questions they answer. Preparation for the conduct as well as the content of the interview is vital by the researcher, for example, the researcher must know the policies very fully and exactly, and not be intimidated by the power of the interviewee (Walford, 2012, p. 113). Further, powerful people, like other interviewees, may not answer questions fully; they may talk blandly or off the point, i.e. with their own agendas, as this may be typical of their usual, often required practice in office (Walford, 2012, p. 113), so the researcher has to ensure that they keep the interview on track, i.e. on their (the researcher's) agenda.

There are difficulties in reporting sensitive research with the powerful, as charges of bias may be difficult to avoid, not least because research reports and publications are placed in the public domain. Walford (2001, p. 141) indicates the risk of libel actions if public figures are named. He asks (1994b, p. 84), 'to what extent is it right to allow others to believe that you agree with them' even if you do not? Should the researcher's own political, ideological or religious views be declared? As Mickelson (1994, p. 147) states: 'I was not completely candid when I interviewed these powerful people. I am far more genuine and candid when I am interviewing non-powerful people'. Deem (1994, p. 156) reports that she and her co-researcher encountered 'resistance and access problems in relation to our assumed ideological opposition to Conservative government education reforms', where access might be blocked 'on the grounds that ours was not a neutral study'.

Mickelson (1994, p. 147) takes this further in identifying an ethical dilemma when 'at times, the powerful have uttered abhorrent comments in the course of the interview'. Should the researcher say nothing, thereby tacitly condoning the speaker's comments, or speak out, thereby risking closing the interview? She contends that, in retrospect, she wished that she had challenged these views and been more assertive (p. 148). She believes that the researcher should challenge different viewpoints, if necessary confrontationally, but this is a high-risk strategy, as the powerful person may simply terminate the interview. Walford (2001) reports the example of an interview with a church minister whose views included ones with which he disagreed:

> AIDS is basically a homosexual disease … and is doing a very effective job of ridding the population of undesirables. In Africa it's basically a non-existent disease in many places.… If you're a woolly woofter, you get what you deserve.… I would never employ a homosexual to teach at my school.
>
> (p. 137)

In researching powerful people Mickelson (1994, p. 132) observes that they are seldom women, yet researchers are often women. This gender divide might prove problematic. Deem (1994, p. 157) reports that, as a woman, she encountered greater difficulty in conducting research than did her male colleague, even though, in fact, she held a more senior position than him. On the other hand, she reports that males tended to be more open with female than male researchers, as female researchers were regarded as less important. Gewirtz and Ozga (1994) report that

> we felt [as researchers] that we were viewed as women in very stereotypical ways, which included being seen as receptive and supportive, and that we were obliged to collude, to a degree, with that version of ourselves because it was productive of the project.
>
> (p. 196)

Walford (2012) notes that, in reality, researching powerful people, approached for whom they are or for the positions that they hold or have held (p. 114), is little different from researching any other people,

except that access may be more problematic, and gaining reliable data may be more challenging. This also means that, unlike other research participants, it is unlikely that anonymity can be offered, indeed the powerful person may insist on being identified.

In approaching researching powerful people, then, it is wise to consider several issues. These are set out in Box 13.3.

13.7 Researching powerless and vulnerable people

Researching powerless people is also a sensitive matter, not least, as Munro *et al*. (2004, p. 299) point out, it is important not to add to their powerlessness. This also applies to vulnerable people: those who are unable to protect their own interests and who may suffer from negative labelling, stigmatization, exclusion or discrimination. (The great claim of participatory research is that it empowers otherwise powerless groups (Healy, 2001; see also Chapter 3).) Powerless people are easily negatively stereotyped and stigmatized (Fiske, 1993; Munro *et al*., 2004), for example: the poor, the unemployed, the homeless, travellers, the disabled, the psychologically disturbed, those with learning difficulties, minority groups, non-heterosexuals, females (Skelton *et al*., 2006) etc.

In conducting research it is important not to add to the disempowerment of already disempowered groups; indeed it may be important actively to promote their empowerment or not to leave them in the condition in which contact was first made (Munro *et al*., 2004, p. 299). (Hammersley (2002, 2014) explores this issue of 'partisan research'; see Chapter 3.)

What does the researcher do, for example, if she finds that women are 'talking down' their own achievements, lives, capabilities or career prospects, such that they will not achieve? If she simply notes this and reports it then she could be seen as complicit in the oppression of women; if she decides not to report it then she could be seen as distorting the research; if she decides to challenge it with the women in question then she could be seen as coming out of the role of the neutral researcher and invading the research site, or indeed to be raising expectations that are not realistic (see also Chapter 3).

Powerless groups may well feel resentful of the well-dressed researcher (Munro *et al*., 2004), even if the researcher's intentions are honourable, or they may feel unable to disclose their true feelings and opinions for fear of bringing yet further negativity to their own situation. They may feel antagonized if interviews are conducted in well-kept surroundings which are very different from their own. Indeed for many, an interview may be the first occasion in their lives that they have experienced such an activity.

Children may well feel powerless and insecure in the presence of a researcher (Greig and Taylor, 1999) and may say what they feel the researcher wishes to hear, what is the school's view, what is socially desirable (p. 131). They may be too shy or embarrassed to reveal their true feelings or to say what really happened in a situation (e.g. child abuse). The researcher must be acutely sensitive to this, and must recognize her/his

BOX 13.3 RESEARCHING POWERFUL PEOPLE

- What renders the research sensitive?
- How to gain and sustain access to powerful people.
- How much are the participants likely to disclose or withhold?
- What is on and off the record?
- How to prepare for interviews with powerful people.
- How to probe and challenge powerful people.
- How, and whether to gain informed consent.
- Is the research overt or covert, with or without deceit?
- How to conduct interviews that balance the interviewer's agenda and the interviewee's agenda and frame of reference.
- How to reveal the researcher's own knowledge, preparation and understanding of the key issues.
- The status of the researcher vis-à-vis the participants.
- Who should conduct interviews with powerful people?
- How neutral and accepting the researcher should be with the participant.
- Whether to identify the participants in the reporting.
- How to balance the public's right to know and the individual's right to privacy.
- What is in the public interest?

own limitations in conducting such research on sensitive matters with vulnerable participants, if necessary handing over such interviews (and, for example, handling projection or displacement techniques) to trained professionals.

The setting for such interviews should be familiar to the children, non-threatening and designed to put them at their ease, to make the strange familiar (Morrison, 2013a), an inversion of Blumer's famous dictum of 'making the familiar strange'. Morrison (2013a) reports on the process of interviewing children (aged 8–9) in a constrained setting in which they were urged to attend interviews in their own out-of-school time and with relative strangers. The interviews were conducted to gather their opinions about a major school innovation brought in by the senior staff of the school and which was evaluated by university staff. Strong asymmetries of power and age were operating in the interviews. Here the interview situation was sensitive in many different ways, and many steps were taken to render them less sensitive and less threatening, indeed enjoyable for the children (discussed in Chapters 14 and 25).

Researchers can conduct honest, sympathetic research on the participants' home ground (as did researchers examining poverty in Hong Kong, who conducted structured interviews in the participants' own homes (Sequeira *et al*., 1996)). They must take care to avoid sounding condescending, patronizing, powerful, domineering or high-handed. This concerns non-verbal behaviour, dress and choice of language (such that it becomes inclusive rather than exclusive, yet without being contrived or artificial). As mentioned in Chapter 7, data are gifts, not entitlements. The researcher has to conduct the research with respect, affording dignity to the participants, whilst not necessarily making promises which cannot be kept (e.g. to change their situation).

The researcher studying powerless and vulnerable groups should be inclusive (i.e. to enable all members of the group in question to participate on an equal footing and to feel valued), and to abide by the ethical principles outlined in Chapter 7 (e.g. informed consent, privacy and confidentiality, recognition of participants' time and efforts, consultation, keeping participants informed, maintaining and concluding relationships, addressing their well-being, indicating any possible adverse effects of participation, ensuring the safety and well-being of researchers) (Connolly, 2003). Powerless participants might feel 'used' in educational research, not only providing data but advancing the careers of the researchers whilst leaving themselves disempowered (see Chapter 7 on 'rape research'). The researcher must avoid this.

Box 13.4 summarizes some key issues in researching powerless and vulnerable people.

BOX 13.4 RESEARCHING POWERLESS AND VULNERABLE GROUPS

- What renders the research sensitive?
- How to gain and sustain access to powerless and vulnerable people.
- How much are the participants likely to disclose or withhold?
- What is on and off the record?
- How to prepare for interviews with powerless and vulnerable people.
- Where will the interviews/data collection take place?
- How to probe powerless and vulnerable people.
- How to ensure non-maleficence and beneficence, dignity and respect.
- How to avoid further stigmatization, negative stereotyping, and marginalization of participants.
- How to act in the interests of the participants.
- How, and whether, to gain informed consent.
- Is the research overt or covert, with or without deceit?
- How to conduct interviews that balance the interviewer's agenda and the interviewee's agenda and frame of reference.
- How to reveal the researcher's own knowledge, preparation and understanding of the key issues.
- How to equalize status between the researcher and the participants.
- How to ensure inclusiveness of participants.
- Who should conduct interviews with powerless and vulnerable people?
- Does the researcher have the expertise to conduct interviews with the participants?
- What protections are there for the participants?
- Whether to identify the participants in the reporting.
- How to balance the public's right to know and the individual's right to privacy.
- What is in the public interest?

Many of the issues raised in considering researching powerful groups are identical to those raised in researching powerless and vulnerable groups (Boxes 13.3 and 13.4). This is deliberate, as both concern ethical, sensitive behaviour, and, though perhaps interpreted differently for the two groups, they apply equally powerfully to both. The Joseph Rowntree Foundation publishes ethical guidelines for researchers working with vulnerable, marginalized groups, powerless people and children.

13.8 Asking questions

Even though an anonymized questionnaire may give participants the freedom to respond in private, in depth and with honesty, and even though a face-to-face interview may be very threatening in connection with some sensitive issues, such that honest or complete answers may be unlikely, as a general rule, the more sensitive the research, the more important it is to conduct face-to-face interviews for data collection. In asking questions in research, Sudman and Bradburn (1982, pp. 50–1) suggest that open questions may be preferable to closed questions and long questions may be preferable to short questions. Both of these enable respondents to answer in their own words, which might be more suitable for sensitive topics. Indeed they suggest that whilst short questions may be useful for gathering information about attitudes, longer questions are more suitable for asking questions about behaviour, and can include examples to which respondents may wish to respond. Longer questions may reduce the under-reporting of the frequency of behaviour addressed in sensitive topics (e.g. the use of alcohol or medication by stressed teachers). On the other hand, the researcher has to be cautious to avoid tiring, emotionally exhausting or stressing the participant by a long questionnaire or interview.

Lee (1993, p. 78) advocates using familiar words in questions as these can reduce a sense of threat in addressing sensitive matters and help the respondent to feel more relaxed. He also suggests the use of 'vignettes' (p. 79): short portrayals of people or situations which contain what are considered to be the important or key factors which affect those people's judgements, decisions or behaviours (p. 79); scenes or short stories about situations or people that can be composed in picture, video, written or spoken formats (Hurworth, 2012, p. 179). These can be part of an interview.

Simon and Tierney (2011) and Hurworth (2012) note that vignettes may be useful in sensitive educational research such as bullying, abuse, assessment, mental health, moral and ethical dilemmas, as participants in the research can give their own reactions to, and accounts of, the positions that they take. They enable the researcher to ask questions about participants' reactions to the situation portrayed, what they would do next or what others might do next. Focusing the discussion away from the individual participant and onto the vignette can 'take the heat out of' the sensitive situation being proposed, i.e. depersonalize it (Hurworth, 2012, p. 179) and reduce the likelihood of receiving only socially desirable or defensive responses by making the sensitivity of the research more unobtrusive (Simon and Tierney, 2011). For example S. Martin (2012, 2013, 2015) shows how this might be undertaken in virtual worlds when exploring sensitive issues of citizenship.

Simon and Tierney (2011) and Hurworth (2012) suggest that vignettes should comprise:

- quite short situations and scenarios that are not only close to the research topic but are rooted in everyday real life or that take real-life examples;
- situations that are credible;
- ordinary everyday situations with which the research participants can connect straightforwardly;
- engaging and interesting age-appropriate and language-appropriate situations which strike a balance between overload of detail (and its resultant complexity) and providing sufficient detail to be interesting;
- deliberately incomplete situations so that there is the potential to enable participants to expand on the situation portrayed;
- characters and events that are relevant and interesting to the participants.

Simon and Tierney (2011) also note that it is important to pilot these for suitability (widely defined) before using them in the research. Vignettes can not only encapsulate concretely the issues under study, but can also deflect attention away from personal sensitivities by projecting them onto another external object – the case or vignette – and the respondent can be asked to react to them personally, for example, 'what would you do in this situation?'.

Researchers investigating sensitive topics have to be acutely percipient of the situation themselves. For example, their non-verbal communication may be critical in interviews. They must, therefore, give no hint of judgement, support or condemnation. They must avoid counter-transference (projecting the researchers' own views, values, attitudes, biases, background onto the situation). Interviewer effects are discussed in Chapter 25 in connection with sensitive research, for example:

- the characteristics of the researcher (e.g. sex, race, age, status, clothing, appearance, rapport, background, expertise, institutional affiliation, political affiliation, type of employment or vocation, e.g. a priest). Females may feel more comfortable being interviewed by a female; males may feel uncomfortable being interviewed by a female; powerful people may feel insulted by being interviewed by a lowly, novice research assistant;
- the expectations that the interviewers may have of the interview (Lee, 1993, p. 99). For example, a researcher may feel apprehensive about, or uncomfortable with, an interview about a sensitive matter. Bradburn and Sudman (1979, in Lee, 1993, p. 101) report that interviewers who did not anticipate difficulties in the interview achieved a 5–30 per cent higher level of reporting on sensitive topics than those who anticipated difficulties. This suggests the need for interviewer training.

Lee (1993, pp. 102–14) suggests several issues in conducting sensitive interviews:

- How to approach the topic (in order to prevent participants' inhibitions and to help them address the issue in their preferred way). Here the advice is to let the topic 'emerge gradually over the course of the interview' (p. 103) and to establish trust and informed consent.
- How to deal with contradictions, complexities and emotions (which may require training and supervision of interviewers); how to adopt an accepting and non-judgemental stance, how to handle respondents who may not be people whom interviewers particularly like or with whom they agree.
- How to handle the operation of power and control in the interview: (a) where differences of power and status operate: where the interviewer has greater or lesser status than the respondent and where there is equal status between the interviewer and the respondent; (b) how to handle the situation in which the interviewer wants information but is in no position to command that this be given and where the respondent may or may not wish to disclose information; (c) how to handle a situation wherein powerful people use the interview as an opportunity for lengthy and perhaps irrelevant self-indulgence; (d) how to handle the situation in which the interviewer, by the end of the session, has information that is sensitive and could give the interviewer power over the respondent and make the respondent feel vulnerable; (e) what the interviewer should do with information that may act against the interests of the people who gave it (e.g. if some groups in society say that they are not clever enough to handle higher or further education); and (f) how to conduct the interview (e.g. conversational, formal, highly structured, highly directed).
- Handling the conditions under which the exchange takes place (Lee, 1993, p. 112) suggests that interviews on sensitive matters should 'have a one-off character', i.e. the respondent should feel that the interviewer and the interviewee may never meet again. This can secure trust, and can lead to greater disclosure than in a situation where a closer relationship between interviewer and interviewee exists. On the other hand, this does not support the development of a collaborative research relationship (Lee, 1993, p. 113).

Much educational research is more or less sensitive; it is for the researcher to decide how to approach the issue of sensitivities and how to address their many forms, allegiances, ethics, access, politics and consequences.

13.9 Conclusion

Educational research is far from a neat, clean, tidy, unproblematic and neutral process; it is shot through with actual and potential sensitivities. With this in mind we have resisted the temptation to provide an exhaustive list of sensitive topics, as this could be simplistic and overlook the fundamental issue which is that it is the *social and individual context* of the research that makes the research sensitive. What may appear to the researcher to be a bland and neutral study can raise deep sensitivities in the minds of the participants. We have argued that it is these that often render the research sensitive rather than, or as well as, the selection of topics of focus. Researchers have to consider the likely or possible effects of the research project, conduct, outcomes, reporting and dissemination not only on themselves but on the participants, on those connected to the participants and on those affected by, or with a stakeholder interest in, the research (i.e. 'consequential validity': the effects of the research). This suggests that it is wise to be cautious and to regard all educational research as potentially sensitive. There are several questions that can be asked by researchers, in their planning, conduct, reporting and dissemination of their studies, and we present these in Box 13.5.

These questions reinforce the importance of regarding ethics as 'situated' (Simons and Usher, 2000), i.e. contingent on particular situations. In this respect sensitive educational research is like any other research, but

BOX 13.5 KEY QUESTIONS IN CONSIDERING SENSITIVE EDUCATIONAL RESEARCH

- What renders the research sensitive?
- What are the obligations of the researcher, to whom, and how will these be addressed? How do these obligations manifest themselves?
- What is the likely effect of this research (at all stages) to be on participants (individuals and groups), stakeholders, the researcher, the community? Who will be affected by the research, and how?
- Who is being discussed and addressed in the research?
- What rights of reply and control do participants have in the research?
- What are the ethical issues that are rendered more acute in the research?
- Over what matters in the planning, focus, conduct, sampling, instrumentation, methodology, reliability, analysis, reporting and dissemination might the researcher have to compromise in order to effect the research? On what can there be compromise? On what can there be no compromise?
- What securities, protections (and from what), liabilities and indemnifications are there in the research, and for whom? How can these be addressed?
- Who is the research for? Who are the beneficiaries of the research? Who are the winners and losers in the research (and about what issues)?
- What are the risks and benefits of the research, and for whom? What will the research 'deliver' and do?
- Should the researcher declare his/her own values, and challenge those with which he/she disagrees or considers to be abhorrent?
- What might be the consequences, repercussions and backlash from the research, and for whom?
- What sanctions might there be in connection with the research?
- What has to be secured in a contractual agreement, and what is deliberately left out?
- What guarantees must and should the researcher give to the participants?
- What procedures for monitoring and accountability must there be in the research?
- What must and must not, should and should not, may or may not, could or could not be disclosed in the research?
- Should the research be covert, overt, partially overt, partially covert, honest in its disclosure of intentions?
- Should participants be identifiable and identified? What if identification is unavoidable?
- How will access and sampling be secured and secure respectively?
- How will access be sustained over time?
- Who are the gatekeepers and how reliable are they?

sharper in the criticality of ethical issues. Also, behind many of these questions of sensitivity lurks the nagging issue of power: who has it, who does not, how it circulates around research situations (and with what consequences) and how it should be addressed. Sensitive educational research is often as much a power play as it is substantive. We advise researchers to regard educational research as involving sensitivities which need to be identified and addressed.

Notes

1 See also Walford (2001, p. 38) in his discussion of gaining access to public schools in the UK, where an early question that was put to him was, 'are you one of us?'.

2 Walford (2001, p. 69) comments on the very negative attitudes of teachers to research on independent schools in the UK, the teachers feeling that researchers had been dishonest and had tricked them, looking only for salacious, sensational and negative data on the school (e.g. on bullying, drinking, drugs, gambling and homosexuality).

Companion Website

The companion website to the book provides PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Validity and reliability

CHAPTER 14

This chapter discusses validity and reliability in educational research. It suggests that both of these terms can be applied to these different types of research, though how validity and reliability are applied to different approaches varies. The chapter proceeds in several stages:

- defining validity
- validity in quantitative, qualitative and mixed methods research
- types of validity
- triangulation
- ensuring validity
- reliability
- reliability in quantitative and qualitative research
- validity and reliability in interviews, experiments, questionnaires, observations, tests, life histories and case studies

There are many different types of validity and reliability. Threats to validity and reliability can never be erased completely; rather the effects of these threats can be attenuated by attention to validity and reliability throughout the research.

Reliability is a necessary but insufficient condition for validity in research; it is a necessary precondition of validity. Brock-Utne (1996, p. 612) contends that the widely held view that reliability is the sole preserve of quantitative research must be exploded, and this chapter demonstrates the significance of her view.

Validity and reliability have different meanings in quantitative, qualitative and mixed methods research. It is important not only to indicate these clearly, but to demonstrate fidelity to the approach in which the researcher is working and to abide by the required principles of validity and reliability. We address this here, locating different interpretations of validity and reliability within different paradigms. One of the purposes of the opening three chapters of this book was to indicate the multiplicity of paradigms. Hence our reference to quantitative and qualitative paradigms here is for simple, heuristic purposes to gain some leverage on the matters involved.

14.1 Defining validity

Validity is an important key to effective research. If a piece of research is invalid then it is worthless. Addressing validity concerns the nature of what is valid, what validity means, how to know if one has achieved an acceptable level of validity, how to address validity in research terms and how validity enters design, inferences and conclusions.

Some versions of validity regard it as essentially a demonstration that a particular instrument in fact measures what it intends, purports or claims to measure, that an account accurately represents 'those features that it is intended to describe, explain or theorise' (Winter, 2000, p. 1).

Other definitions state that validity is the extent to which interpretations of data are warranted by the theories and evidence used (Ary *et al*., 2002, p. 267). The issue of warrants was explored in Chapter 11, arguing that researchers must indicate the grounds and the evidence that they will use to connect their data with the claims made from, or conclusions drawn from, the data. A warrant, as Chapter 11 noted, is the logical link made between data and proposition, between data and conclusions (Andrews, 2003, p. 30), which supports the weight given to the explanation offered in the face of alternative, rival explanations. We advise the reader to review the discussion of warrants in Chapter 11. A piece of research is valid if the warrants that underpin it are defensible and, thereby, if the conclusions drawn and the explanations given can stand their ground in the face of rival conclusions and explanations; validity and warrants are linked intimately.

As researchers, we must be certain that our instruments for understanding phenomena are as sound as possible, i.e. that they are valid. This is particularly the case for abstract, unclearly or indirectly observable, theoretical constructs such as 'intelligence', 'creativity', 'anxiety', 'motivation', 'extraversion' and 'empathy', for which no natural measures or units of measurement exist (cf. Shadish *et al*., 2002, p. 65). How can we be sure that our instruments for gathering data on these unseen, theoretical constructs are safe and

that the proxies we use to assess them are valid? How can we be sure that the observable tasks and features that we choose are fair representations and indicators of these abstract concepts? How can we defensibly construct, name and define an abstract concept, and how do we know that a particular construct is prototypical or socio-culturally and contextually bound (pp. 66–7)? This raises the issue of construct validity, and we address this important factor in this chapter.

In qualitative research, given that multiple views of 'reality' exist, whose is credible and 'correct', how do we know and how do we validate socially constructed knowledge (Flick, 2009, p. 389)? Ary *et al*. (2002) note that validity not only concerns the extent to which an instrument measures what it claims to measure, but that the meaning and interpretation of the results of the data collection and instrumentation are sound (p. 242).

This chapter, in discussing the limits of discourses on validity, argues for a need to move beyond technical issues of how to address it and to address the ontological and epistemological natures (plural) of validity. We engage these issues as well as how researchers can address and ensure validity.

Shadish *et al*. (2002, pp. 37–8) identify four main kinds of validity:

- construct validity: the validity of inferences made about the nature and manifestations of theoretical factors;
- statistical conclusion validity: the use of appropriate statistics to determine, for example, correlation between intervention and outcome;
- internal validity: the validity of inferred and found relationships between elements of the research design and outcomes;
- external validity: generalizability.

They note that both construct validity and external validity concern generalization: the former with regard to the derivation and operation of theoretical constructs, and the latter with regard to sampling. There are, however, several different kinds of validity which fall into the four categories above, for example:

- catalytic validity;
- concurrent validity;
- consequential validity;
- construct validity;
- content validity;
- criterion-related validity;
- convergent and discriminant validity;
- cross-cultural validity;
- cultural validity;
- descriptive validity;
- ecological validity;
- evaluative validity;
- external validity;
- face validity;
- internal validity;
- interpretive validity;
- jury validity;
- predictive validity;
- statistical conclusion validity;
- systemic validity;
- theoretical validity.

It is not our intention in this chapter to discuss all of these terms in depth. Rather the main types of validity will be addressed. The argument will be made that, whilst some of these terms are more comfortably the preserve of quantitative methodologies, this is not exclusively the case. Indeed validity is the touchstone of all types of educational research. Hence the researcher will need to locate her discussions of validity within the research paradigm that is being used. This is not to suggest, however, that research should be paradigm-bound, that is a recipe for stagnation and conservatism; rather validity should be fit for purpose.

Validity takes many forms. For example, in qualitative data validity might be addressed through the honesty, depth, authenticity, richness, trustworthiness, dependability, credibility and scope of the data achieved, the participants approached, the extent of triangulation and the disinterestedness or objectivity of the researcher (Winter, 2000; Flick, 2009). This also means that the matters reported, for example, in an interview, are correct, 'socially appropriate' (Flick, 2009, p. 388) and given sincerely, echoing Habermas's (1979, 1982) views introduced in Chapter 3 of the need for a communication to be true, sincere, legitimate, truthfully given and comprehensible. We pick up this point below, in discussions of mixed methods research.

It is impossible for research to be 100 per cent valid; that is the optimism of perfection. Validity should be seen as a matter of degree rather than as an absolute state (Gronlund, 1981). Hence at best we strive to minimize invalidity and maximize validity.

14.2 Validity in quantitative research

In much quantitative research, validity often (not always) strives to be faithful to several features, for example:

- controllability;
- replicability;

- consistency;
- predictability;
- the derivation of generalizable statements of behaviour;
- randomization of samples;
- neutrality/objectivity;
- observability.

In many cases validity involves being faithful to the assumptions underpinning the statistics used, the construct and content validity of the measures used, careful sampling and the avoidance of a range of threats to internal and external validity outlined later in this chapter.

Statistical conclusion validity (Shadish *et al*., 2002) may be threatened by, for example: low statistical power; violating assumptions in the statistics used (e.g. of normal distributions of data, of linearity, of sample size); measurement error; too limited a range in the data derived from the measures used; too much variation in the procedures for the treatments/interventions in question; extraneous variables (e.g. moderator and mediator variables); wide variability in the outcome measures; built-in error in the statistics used (e.g. their formulae); a false assumption of causality.

14.3 Validity in qualitative research

Much qualitative research abides by principles of validity which differ in many respects from those of quantitative methods. Validity in qualitative research has several principles (Lincoln and Guba, 1985; Bogdan and Biklen, 1992; Ary *et al*., 2002; Flick, 2009):

- the natural setting is the principal source of data;
- context-boundedness and 'thick description';
- data are socially situated, and socially and culturally saturated;
- the researcher is part of the researched world;
- as we live in an already interpreted world, a doubly hermeneutic exercise (Giddens, 1979) is necessary to understand others' understandings of the world; the paradox here is that the most sufficiently complex instrument to understand human life is another human (Lave and Kvale, 1995, p. 220), but this risks human error in all its forms;
- holism in the research;
- the researcher – rather than a research tool – is the key instrument of research;
- data are descriptive;
- there is a concern for processes rather than solely with outcomes;
- data are analysed inductively rather than using a priori categories;
- data are presented in terms of the respondents rather than the researcher;
- seeing and reporting the situation through the eyes of participants (Geertz, 1974);
- respondent validation is important;
- catching agency, meaning and intention are essential.

Maxwell (1992) argues that qualitative researchers should avoid working within the agenda of the positivists in arguing for the need for research to demonstrate concurrent, predictive, convergent, criterion-related, internal and external validity. However, the discussion below indicates that this need not be so. Guba and Lincoln (1989) argue for the need to replace positivist notions of validity in qualitative research with 'authenticity'. Maxwell (1992), echoing Mishler (1990), suggests that 'understanding' is a more suitable term than 'validity' in qualitative research. We, as researchers, are part of the world that we are researching, and we cannot be completely objective about that, hence other people's perspectives are equally as valid as our own, and the task of research is to uncover these. Validity, then, concerns the meanings that subjects give to data and inferences drawn from the data (Hammersley and Atkinson, 1983). 'Fidelity' (Blumenfeld-Jones, 1995) requires the researcher to be as honest as possible to the self-reporting of the researched.

Agar (1993) notes that, in qualitative data collection, the intensive personal involvement and in-depth responses of individuals secure a sufficient level of validity and reliability. This claim is contested by Hammersley (1992, p. 144, 2011) and Silverman (1993, p. 153), who argue that these are insufficient grounds for validity and reliability, and that the individuals concerned have no privileged position on interpretation. (Of course, neither are actors 'cultural dopes' who need a sociologist or researcher to tell them what is 'really' happening.) Silverman argues that, whilst immediacy and authenticity make for interesting journalism, ethnography must have different but equally rigorous notions of validity and reliability. This involves moving beyond selecting data simply to fit a preconceived or ideal conception of the phenomenon or because they are spectacularly interesting (Fielding and Fielding, 1986). Data selected must be representative of the sample, the whole data set and the field, i.e. they must address content, construct and concurrent validity.

Hammersley (1992, pp. 50–1, 2011) suggests that validity in qualitative research replaces certainty with confidence in our results, and that, as reality is independent of the claims made for it by researchers, our accounts will only be representations of that reality

rather than reproductions of it. Lincoln and Guba (1985) and Ary *et al*. (2002) suggest that key criteria of validity in qualitative research are:

- credibility: the truth value (replacing the quantitative concepts of internal validity);
- transferability: generalizability (replacing the quantitative concept of external validity);
- dependability: consistency (replacing the quantitative concept of reliability);
- confirmability: neutrality (replacing the quantitative concept of objectivity).

Lincoln and Guba (1985) argue that, within these criteria of validity, rigour can be achieved by careful audit trails of evidence, member checking/respondent validation (confirmation by participants) when coding or categorizing results, peer debriefing, negative case analysis, 'structural corroboration' (triangulation, discussed below) and 'referential material adequacy' (adequate reference to standard materials in the field). Trustworthiness, they suggest, can be addressed in the credibility, fittingness, auditability and confirmability of the data (see also Morse *et al*., 2002).

Whereas quantitative data place great store on both external validity and internal validity, the emphasis in much qualitative research is on internal validity, and in many cases external validity is an irrelevance for qualitative research (Winter, 2000, p. 8; Creswell, 2012) as it does not seek to generalize but only to represent the phenomenon being investigated, fairly and fully. Of course, some qualitative research, for example, Miles and Huberman (1994), does move towards generalizability, and indeed Chapter 2 indicates that qualitative data can be 'quantitized'. The overwhelming feature of qualitative research is its concern with the phenomenon or situation in question, and not generalizability (Hammersley, 2013). Hence issues such as random sampling, replicability, alpha coefficients of reliability, isolation and control of variables, and predictability simply do not matter in much qualitative research.

Maxwell (1992) argues for five kinds of validity as 'understanding' in qualitative methods:

- descriptive validity: the factual accuracy of the account, that it is not made up, selective or distorted (cf. Winter, 2000, p. 4); in this respect validity subsumes reliability; it is akin to Blumenfeld-Jones's (1995) notion of 'truth' in research – what actually happened (objectively factual) – and to Glaser's and Strauss's (1967) term 'credibility';
- interpretive validity: the ability of the research to catch the meaning, interpretations, terms and intentions that situations and events, i.e. data, have for the participants/subjects themselves, in their terms; it is akin to Blumenfeld-Jones's (1995) notion of 'fidelity' – what it means to the researched person or group (subjectively meaningful); interpretive validity has no clear counterpart in experimental/positivist methodologies;
- theoretical validity: the theoretical constructions that the researcher brings to the research (including those of the researched); theory here is regarded as explanation; theoretical validity is the extent to which the research explains phenomena; in this respect it is akin to construct validity (discussed below); in theoretical validity the constructs are those of all the participants;
- generalizability: the view that the theory generated may be useful in understanding other similar situations; generalizing here refers to generalizing within specific groups or communities, situations or circumstances validly, and, beyond, to specific outsider communities, situations or circumstance (external validity);
- evaluative validity: the application of an evaluative, judgemental stance towards that which is being researched, rather than a descriptive, explanatory or interpretive framework.

To these one can add Auerbach's and Silverstein's (2003) category of *transparency*, i.e. how far the reader can understand, and is informed of, the processes by which the interpretation made is actually reached (cf. Teusner, 2016). Indeed Teusner (2016), commenting on insider research, argues that by making the procedures of the research transparent, with results and conclusions demonstrating clarity and justifiability (rehearsing the comments below and Chapter 11 on 'warrants'), this renders external validation less important (p. 88).

Central to Teusner's views of transparency in insider research is the importance of reflexivity and disclosure; she argues for researchers to address concerns about: (a) whether the relationship between the researcher and participants has a negative impact on the participants' behaviour; (b) whether the researcher's tacit knowledge will risk misinterpreting data, making false assumptions or missing potentially important information; (c) whether the researcher's own politics, loyalties, perspectives, socio-cultural and moral standpoints and agendas will lead to misrepresentation or distortion; (d) whether the researcher's own emotional connections with participants will impact on the research; and (e) how far the researcher's and participants' status will impact on the research relationships (2016, pp. 90–4).

Validity in qualitative research concerns the purposes of the participants, the actors and the appropriateness of the data-collection methods used to catch those purposes (Winter, 2000, p. 7). Maxwell (2005) suggests that validity here can be enhanced by 'intensive long-term involvement', 'rich data', 'respondent validation', 'intervention' (e.g. in action research or case study research), 'searching for discrepant evidence and negative cases', 'triangulation' and 'comparison' (e.g. between a control group and an intervention group, or between groups in different sites and location) (pp. 110–14) and by considering alternative explanations of a phenomenon (p. 126).

Differences in the meanings and criteria for validity in quantitative and qualitative are summarized in Table 14.1.

Clearly the criteria are not the exclusive preserve of each of the two main types of research here (quantitative and qualitative). The intention of Table 14.1 is heuristic and to indicate emphases only.

Onwuegbuzie and Leech (2006b, pp. 239–46) set out many steps that researchers can take to ensure validity in qualitative research (several of which derive from Lincoln and Guba, 1985; see also Huberman and Miles, 1998; Ary *et al*., 2002; Teddlie and Tashakkori, 2009, pp. 295–7; Flick, 2009; Yin, 2009; Teusner, 2016). These include:

- prolonged engagement in the field (to gather rich and sufficient data);
- persistent observation (to identify key relevant issues and to separate these from comparative irrelevancies);
- triangulation (discussed later in this chapter: data, perspectives, instruments, time, methodologies, people etc.): 'structural corroboration';
- leaving an audit trail (documentation and records used in the study that include: raw data; records of analysis and data reduction; reconstructions and syntheses of data; 'process notes' (on how the research and analysis are proceeding; notes on 'intentions and dispositions' of the researcher as the study proceeds; information concerning the development of instruments for data collection));
- member checking/informant feedback (respondent validation, discussed below);
- weighting the evidence, ensuring that correct attention is paid to higher-quality data (e.g. those data gathered from long engagement, detailed study and trusted participants) and less attention is paid to low-quality data;
- checking for representativeness (ensuring that unsupported generalizability of the findings is avoided);
- checking for researcher effects/clarifying researcher bias (how far the personal biases, assumptions or values of the researcher, or how far the researcher's personal characteristics (e.g. clothing, appearance, sex, age, ethnicity) affect the research), premature closure of data collection, unexplored data which are contained in field notes and too close an empathy between researcher and subjects;

TABLE 14.1 COMPARING VALIDITY IN QUANTITATIVE AND QUALITATIVE RESEARCH

Bases of validity in quantitative research		*Bases of validity in qualitative research*
Controllability	⟷	Natural
Isolation, control and manipulation of required variables	⟷	Thick description and high detail on required or important aspects
Replicability	⟷	Uniqueness
Predictability	⟷	Emergence, unpredictability
Generalizability	⟷	Uniqueness
Context-freedom	⟷	Context-boundedness
Fragmentation and atomization of research	⟷	Holism
Randomization of samples	⟷	Purposive sample/no sampling
Neutrality	⟷	Value-ladenness of observations/double hermeneutic
Objectivity	⟷	Confirmability
Observability	⟷	Observability and non-observable meanings and intentions
Inference	⟷	Description, inference and explanation
'Etic' research	⟷	'Emic' research
Internal validity	⟷	Credibility
External validity	⟷	Transferability
Reliability	⟷	Dependability
Observations	⟷	Meanings

- making contrast/comparisons (e.g. between subgroups, sites, literature);
- theoretical sampling (following the data and where they lead, rather than leading the data, and ensuring that the research addresses all the required aspects of the theory);
- checking the meaning of outliers (rather than ignoring outliers and exceptions, researchers should examine them to see what leverage they provide into an understanding of the phenomenon in question);
- using extreme cases (e.g. to identify what is missing in the majority of cases);
- ruling out spurious relations (avoiding attributing causality or association where none exists);
- replicating a finding (identifying how far the findings might apply to other groups);
- referential adequacy (how well-referenced the findings are to benchmark or significant literature);
- following up surprises (avoiding ignoring surprise results);
- structural relationships (looking for consistency between the findings – with each other and with literature);
- peer review;
- peer debriefing (external evaluation of the research, its conduct and findings);
- reflexivity and control of bias;
- rich and thick description (providing detail to support and corroborate findings);
- the 'modus operandi' approach (specifically looking for possible sources of invalidity in the research);
- assessing rival explanations (looking for alternative interpretations and explanations of the data);
- negative case analysis (examining disconfirming cases to see if the hypotheses or findings need to be amended in light of them);
- checking that the findings are thoroughly grounded in data, that inferences made are logical, that strategies for analysis are used correctly and that the category structure is appropriate;
- confirmatory data analysis (conducting qualitative replication studies where possible);
- theoretical adequacy (by, for example, theory triangulation and extended fieldwork);
- effect sizes (avoiding simply 'binarizing' matters (e.g. strong/weak; present/absent; positive/negative) and replacing them with indications of size/power or strength of the findings).

This comprehensive list of ways of striving to ensure validity in qualitative research has similarities in some places with those of quantitative research (e.g. replication, avoidance of researcher bias, external evaluation, representativeness, suitable generalizability, theoretical sampling, triangulation, transparency, etc.). This suggests that, whilst there may be different canons of validity between quantitative and qualitative research, and whilst there may be different interpretations of the meaning of 'validity' in different kinds of research, nevertheless there is some common ground between them; they are not mutually exclusive.

14.4 Validity in mixed methods research

Though each of the methods in mixed methods research (MMR) has to conform to its specific validity requirements in quantitative and qualitative research, there is an argument for identifying specific validity requirements for MMR. Onwuegbuzie and Johnson (2006) argue that the term 'validity' should be replaced by 'legitimation' in MMR, and they identify nine main types of legitimation (discussed below). These nine methods, the authors aver (p. 52), constitute an attempt to overcome problems in MMR of:

- representation (using largely or only words and pictures to catch the dynamics of lived experiences and unfolding, emergent situations);
- legitimation (ensuring that the results are dependable, credible, transferable, plausible, confirmable and trustworthy);
- integration (using and combining quantitative and qualitative methods, each with their own, sometimes antagonistic canons of validity, e.g. quantitative data may use large random samples whilst qualitative data may use small, purposive samples, and yet they may be placed on an equal footing) (p. 54).

Their nine types of legitimation in MMR (Onwuegbuzie and Johnson, 2006, p. 57) are:

1. *Sample integration* (how far different kinds and sizes of sample in combination, or the same samples in quantitative and qualitative research, can enable high-quality inferences to be made).
2. *Inside-outside* (how far researchers use, combine and balance both insiders' views ('*emic*' research) and outsiders' views ('*etic*', objective research) in the research in describing and explaining).
3. *Weakness minimization* (how far any weaknesses that stem from one approach are compensated by the strengths of the other approach, together with suitably weighting such strengths and weaknesses).
4. *Sequential* (how far one can minimize order effects (quantitative to qualitative and vice versa) in

'meta-inferences' made from data collection and analysis, such that one could reverse the order of the inferences made, or the order of the quantitative and qualitative data, without loss of power to the 'meta-inferences').

5 *Conversion* (how far qualitizing numerical data or quantitizing qualitative data can assist in yielding robust 'meta-inferences').

6 *Paradigmatic mixing* (how successful is the combination of the ontological, epistemological, axiological, methodological and rhetorical beliefs and practices in yielding useful results, particularly if the paradigms are in tension with each other).

7 *Commensurability* (how far any 'meta-inferences' made from the data catch a 'mixed worldview' (i.e. rejecting the incommensurability of paradigms) that is enabled by 'Gestalt switching' and integration of paradigms and their methodologies).

8 *Multiple validities* (fidelity to the canons of validity for each of the quantitative and qualitative data gathered).

9 *Political* (how accepted to the audiences are the 'meta-inferences' stemming from the combination of quantitative and qualitative methods).

Collins *et al*. (2012) add two criteria which concern philosophical clarity, researchers' assumptions and connecting quality criteria from different communities involved in MMR:

10 *Holistic legitimation* (the inclusion of major works to demonstrate legitimation and quality); and

11 *Synergistic legitimation*, where combining the process and outcome of legitimation is superior to addressing these two separately; adopting a dialectical process of multiple perspectives, philosophical assumptions and stances; regarding as equally important the legitimation processes in quantitative and qualitative approaches; and balancing opposing quantitative and qualitative approaches (p. 855).

Long (2015), however, argues that discussions of validity in MMR is still at an early stage. Commenting on the work of Collins *et al*. (2012), she advocates taking the issue of validity in MMR wider than is typically found, suggesting that, to date, validity in MMR has been confined to matters of design, procedures, methods and techniques, i.e. 'the logic of justification'. She argues for a broader embrace of validity, to include fundamental issues in the ontology and epistemology of validity in MMR. Here, she draws on Habermas's criteria for speech-act validity claims in communicative action, arguing that validity comprises: sincerity, legitimacy, truthfulness, rightness and comprehensibility in 'action oriented to mutual understanding' (Habermas, 1972, p. 310). In turn, this addresses Habermas's *ideal speech situation* which is 'discursively redeemed' in intersubjective, dialogic speech acts (Habermas, 1979, p. 2, 1984, p. 10; Morrison, 1995a, p. 104). Validity in MMR, thus construed, concerns, for example (Morrison, 1995a, p. 105):

- orientation to a 'common interest ascertained without deception';
- freedom to enter a discourse and to check questionable claims;
- freedom to evaluate explanations and to modify a given conceptual framework;
- freedom to reflect on the nature of knowledge, to assess justifications and to alter norms;
- freedom to allow commands or prohibitions to enter discourse when they can no longer be taken for granted;
- freedom to reflect on the nature of political will;
- mutual understanding between participants;
- equal opportunity to select and employ speech acts and to join a discussion, with that discussion being free from domination and distorting or deforming influences;
- recognition of the legitimacy of each subject to participate in the dialogue as an autonomous and equal partner;
- the consensus resulting from discussion derives from the force of the better argument alone, and not from the positional or political power of the participants;
- all motives except the cooperative search for truth are excluded.

Though Long (2015) draws attention to some challenges in this conception of validity, she understates the critiques of Habermas's view (for an account of these, see Morrison, 1995a).

In the following sections, which describe types of validity, where it is useful to separate the interpretations of validity in quantitative and qualitative research, this has been done. In some cases (e.g. catalytic, consequential validity), as the issues remain the same regardless of the type of research, this separation has not been done. The scene is set by considerations of internal and external validity, and then other types of validity are considered.

14.5 Types of validity

Internal validity

Both qualitative and quantitative methods can address internal and external validity. Internal validity seeks to demonstrate that the explanation of a particular event, issue or set of data which a piece of research provides can actually be sustained by the data and the research (cf. Shadish *et al*., 2002, p. 37). This requires, inter alia, accuracy and correctness, which can be applied to both quantitative and qualitative research. The findings must describe accurately the phenomena being researched. Onwuegbuzie and Leech (2006b, p. 234) define internal validity as the 'truth value, applicability, consistency, neutrality, dependability, and/or credibility of interpretations and conclusions within the underlying setting or group'.

Internal validity in quantitative research

The following summaries adapted from Campbell and Stanley (1963), Bracht and Glass (1968), Lewis-Beck (1993), Shadish *et al*. (2002) and Creswell (2012) distinguish between 'internal validity' and 'external validity'. Internal validity is concerned with the question, do the experimental treatments, in fact, make a difference in the specific experiments under scrutiny? Is the research sufficiently free of errors or violations of validity? Is the research secure? External validity, on the other hand, asks the question, 'given these demonstrable effects, to what populations or settings can they be generalized?'.

There are several kinds of threat to internal validity in quantitative research (many of these apply strongly, though not exclusively, to experimental research), for example:

- *History*: Frequently in educational research, events other than the intervention treatments occur during the time between pre-test and post-test observations (e.g. in a longitudinal survey, experiment, action research). Such events produce effects that can mistakenly be attributed to differences in treatment.
- *Maturation*: Between any two observations, subjects change in a variety of ways. Such changes can produce differences that are independent of the research. The problem of maturation is more acute in protracted educational studies than in brief laboratory experiments.
- *Ambiguous temporal precedence*: It is important to disclose which variable is taken to be the cause and which the effect (the direction of causality).
- *Statistical regression*: Regression means simply that subjects scoring highest on a pre-test are likely to score relatively lower on a post-test; conversely, those scoring lowest on a pre-test are likely to score relatively higher on a post-test. In short, in pre-test–post-test situations, there is regression to the mean. Regression effects can lead the educational researcher mistakenly to attribute post-test gains and losses to low scoring and high scoring respectively. Like maturation effects, regression effects increase systematically with the time interval between pre-tests and post-tests (e.g. in action research, experiments or longitudinal research). Statistical regression occurs in educational research due to the unreliability of measuring instruments and to extraneous factors unique to each group, for example, in an experiment.
- *Testing*: Pre-tests at the beginning of research (e.g. experiments, action research, observational research) can produce effects other than those due to the research treatments. Such effects can include sensitizing subjects to the true purposes of the research and practice effects which produce higher scores on post-test measures.
- *Instrumentation*: Unreliable tests or instruments can introduce serious errors into research (e.g. testing, surveys, experiments). With human observers or judges or changes in instrumentation and calibration, error can result from changes in their skills and levels of concentration over the course of the research.
- *Selection*: Bias may be introduced as a result of differences in the selection of subjects for the comparison groups or when intact classes are employed as experimental or control groups. Selection bias may interact with other factors (history, maturation, etc.) to cloud further the effects of the comparative treatments.
- *Experimental mortality (attrition)*: The loss of subjects through dropout often occurs in long-running research (e.g. experiments, longitudinal research, action research) and may confound the effects of the variables, for whereas initially the groups may have been randomly selected, those who stay the course may be different from the unbiased sample that began it.
- *Instrument reactivity*: The effects that the data-collection instruments exert on the people in the study (e.g. observations, questionnaires, video recordings, interviews).
- *Selection-maturation interaction*: Where there is confusion between the research design effects and the variable's effects.
- *Type I and Type II errors*: A false positive and a false negative, respectively.

A Type I error can be addressed by setting a more rigorous level of significance (e.g. $\rho < 0.01$ rather than $\rho < 0.05$). Boruch (1997, p. 211) suggests that a Type II error may occur if: (a) the measurement of a response to the intervention is insufficiently valid; (b) the measurement of the intervention is insufficiently relevant; (c) the statistical power of the experiment is too low; (d) the wrong population was selected for the intervention. A Type II error can be addressed by reducing the level of significance (e.g. $\rho < 0.20$ or $\rho < 0.30$ rather than $\rho < 0.05$). The more one reduces the chance of a Type I error the more chance there is of committing a Type II error, and vice versa. We discuss Type I and Type II errors in Chapter 39.

Ary *et al*. (2002) suggest that one threat to internal validity stems from 'construct underrepresentation' (p. 243): the under-representation of a construct in instrumentation or data collection (e.g. too narrow, too selective), whilst another threat is from 'construct-irrelevance variance' (p. 243): the effect of other, extraneous factors on the factor or process in question.

Later in this chapter we address how these threats might be mitigated.

Internal validity in qualitative research

In ethnographic, qualitative research there are several main kinds of internal validity (LeCompte and Preissle, 1993, pp. 323–4):

- confidence in the data;
- the authenticity of the data (the ability of the research to report a situation through the eyes of the participants);
- the cogency of the data;
- the soundness of the research design;
- the credibility of the data;
- the auditability of the data;
- the dependability of the data;
- the confirmability of the data.

Writers on the issue of authenticity, argue for:

- fairness (that there should be a complete and balanced representation of the multiple realities in, and constructions of, a situation);
- ontological authenticity (the research should provide a fresh and more sophisticated understanding of a situation, e.g. making the familiar strange (Blumer, 1969), a significant feature in reducing 'cultural blindness' in a researcher, a problem which might be encountered in moving from being a participant to being an observer (Brock-Utne, 1996, p. 610));
- educative authenticity (the research should generate a new appreciation of these understandings);
- catalytic authenticity (the research gives rise to specific courses of action);
- tactical authenticity (the research should bring benefit to all involved: the ethical issue of 'beneficence').

Hammersley (1992, p. 71) suggests that internal validity for qualitative data requires attention to:

- plausibility and credibility;
- the kinds and amounts of evidence required (such that the greater the claim that is being made, the more convincing the evidence has to be for that claim);
- clarity on the kinds of claim made from the research (e.g. definitional, descriptive, explanatory, theory generative).

In ethnographic research internal validity can be addressed by using low-inference descriptors, multiple researchers, participant researchers, peer examination of data and mechanical means to record, store and retrieve data (LeCompte and Preissle, 1993, p. 338). By tracking and storing information clearly, it is possible for the ethnographer to eliminate rival explanations of events and situations.

Lincoln and Guba (1985, pp. 219, 301) suggest that credibility in naturalistic inquiry can be addressed by:

- prolonged engagement in the field;
- persistent observation (in order to establish the relevance of the characteristics for the focus);
- triangulation (of methods, sources, investigators and theories);
- peer debriefing (exposing oneself to a disinterested peer in a manner akin to cross-examination, in order to test honesty, working hypotheses and to identify the next steps in the research);
- negative case analysis (in order to establish a theory that fits every case, revising hypotheses retrospectively);
- member checking (respondent validation to assess intentionality, to correct factual errors, to offer respondents the opportunity to add further information or to put information on record, to provide summaries and to check the adequacy of the analysis).

Whereas in quantitative research, history and maturation are viewed as threats to the validity of the research, ethnographic research simply assumes that this will happen; ethnographic research allows for change over time – it builds it in. Internal validity in ethnographic research is also addressed by the reduction of observer

effects by having the observers sample widely and stay in the situation for such a long time that their presence is taken for granted.

Onwuegbuzie and Leech (2006b, pp. 235–7) identify twelve kinds of threat to internal validity in qualitative research:

1. *Ironic legitimation* (how far the research recognizes and is able to work with multiple realities and interpretations of the same situation, even if they are simultaneously contradictory).
2. *Paralogical legitimation* (how far the research is able to catch and address paradoxes in the claims to validity).
3. *Rhizomatic legitimation* (how much the research loses data when mapping of data rather than describing takes place).
4. *Voluptuous legitimation* (how far the interpretation placed on the data exceeds the capability of the researcher to support that interpretation from the data).
5. *Descriptive validity* (the accuracy of the account given by the researcher).
6. *Observational bias* (inadequate sampling of words, observations or behaviours in the study).
7. *Researcher bias* (discussed earlier).
8. *Reactivity* (how far the research alters the situation being researched or the participants in the research, e.g. the Hawthorne effect (discussed below) and the novelty effect).
9. *Confirmation bias* (the tendency for a piece of research to confirm existing findings or hypotheses).
10. *Illusory confirmation* (the tendency to find relationships, e.g. between people, behaviours or events, when in fact they do not exist).
11. *Causal error* (inferring causal relations when none exists or where no evidence has been provided of their existence).
12. *Effect size* (avoiding taking numerical effect sizes and qualitizing them, when such a step would enrich the analysis; failure to take into account effect sizes and the meaningfulness that they could bring to the interpretation of the data).

Researchers need to be alert to these potential sources of invalidity and take steps to avoid or minimize them.

External validity

External validity refers to the degree to which the results can be generalized to the wider population, cases, settings, times or situations, i.e. to the transferability of the findings. The issue of generalization is problematical. For some researchers generalizability is a *sine qua non*, whilst this is far less the case in other kinds of research (e.g. naturalistic research). For one school of thought, generalizability through stripping out contextual variables is fundamental, whilst, for another, generalizations which say little about the context have little that is useful to say about human behaviour. For positivists and post-positivists, variables must be isolated and controlled and samples randomized, whilst for ethnographers human behaviour is infinitely complex, irreducible, socially situated and unique.

External validity in quantitative research

External validity in quantitative research concerns generalizability: how far we can generalize from a sample to a population. In addressing external validity, attention must be paid to a range of challenges. These include, for example (Morrison, 2001; Shadish *et al*., 2002; Cartwright and Hardie, 2012):

- generalizing from a narrow sample or sub-groups to a broad population;
- generalizing from a sample to an even smaller sample (sub-group or individuals) (the ecological fallacy);
- generalizing from one situation to another similar situation without taking account of contextual and causal differences;
- generalizing from one situation to another dissimilar situation without taking account of differences of context and causal similarities;
- the exception fallacy: deriving a generalized statement on the basis of exceptional cases;
- generalizing from unstandardized, under-controlled variable treatments (e.g. the failure to keep to the same processes or the overlooking of other factors present in the situation);
- overlooking the range of outcomes of an intervention (too tight a focus on certain outcomes, to the neglect of other outcomes); for example, an intervention that puts greater pressure on students' measured performance in mathematics might overlook the negative fallout of this.

Threats to external validity are likely to limit the degree to which generalizations can be made from the particular – for example, experimental – conditions to other populations or settings. Below, we summarize a number of factors that jeopardize external validity (adapted from Campbell and Stanley, 1963; Bracht and Glass, 1968; Hammersley and Atkinson, 1983; Vulliamy, 1990; Lewis-Beck, 1993; Onwuegbuzie and Johnson, 2006; Creswell, 2012; Cartwright and Hardie, 2012).

- *Failure to describe independent variables explicitly*: unless independent variables are adequately described by the researcher, future replications of the research conditions are virtually impossible.
- *Lack of representativeness of available and target populations*: whilst participants in the research may represent an available population, they may not represent the population to which the researcher seeks to generalize her findings, i.e. poor sampling and/or randomization.
- *Hawthorne effect*: medical research has long recognized the psychological effects that arise out of mere participation in drug experiments, and placebos and double-blind designs are commonly employed to counteract the biasing effects of participation. Similarly, so-called Hawthorne effects threaten to contaminate research treatments in educational research when subjects realize their role as guinea pigs.
- *Inadequate operationalizing of dependent variables*: dependent variables that the researcher operationalizes must have validity in the non-research setting to which she wishes to generalize her findings. A questionnaire on career choice, for example, may have little validity in respect of the actual employment decisions made by undergraduates on leaving university.
- *Sensitization/reactivity to experimental/research conditions*: as with threats to internal validity, pretests may cause changes in the subjects' sensitivity to the intervention variables and thus cloud the true effects of the treatment.
- *Interaction effects of extraneous factors and experimental/research treatments*: all of the above threats to external validity represent interactions of various clouding factors with treatments. As well as these, interaction effects may also arise as a result of any or all of those factors in different combinations (see also threats to internal validity).
- *Invalidity or unreliability of instruments*: the use of instruments which yield data in which confidence cannot be placed (see below on tests).
- *Ecological validity, and its partner, the extent to which behaviour observed in one context can be generalized to another*: Hammersley and Atkinson (1983, p. 10) comment on the problems that surround attempts to relate inferences from responses gained under experimental conditions, or from interviews, to everyday life. Cartwright and Hardie (2012) comment in detail on the difficulties in applying the findings from an experiment in one context to a different location.
- *Multiple treatment validity*: applying several treatments simultaneously or in sequence may cause interaction effects between these treatments, such that it is difficult, if not impossible, to isolate the effects of particular treatments.

External validity in qualitative research

Generalizability in naturalistic research is interpreted as comparability and transferability (Lincoln and Guba, 1985; Eisenhart and Howe, 1992, p. 647). These writers suggest that it is possible to assess the typicality of a situation – the participants and settings – to identify possible comparison groups, and to indicate how data might translate into different settings and cultures (see also Strauss and Corbin, 1990; LeCompte and Preissle, 1993, p. 348). Schofield (1996, p. 200) comments that it is important in qualitative research to provide a clear, detailed and in-depth description so that others can decide the extent to which findings from one piece of research are generalizable to another situation, i.e. to address the twin issues of *comparability* and *translatability* (cf. Cartwright and Hardie's (2012) comments on the need for there to be similarly between the causal processes in the locations of the original research and those in other locations).

Qualitative research can be generalizable (Schofield, 1996, p. 209), by studying the typical for its applicability to other situations – the issue of *transferability* (see also LeCompte and Preissle, 1993, p. 324) – and by performing multi-site studies (e.g. Miles and Huberman, 1984), though it could be argued that this is injecting a degree of positivism into non-positivist research. Lincoln and Guba (1985, p. 316) caution the naturalistic researcher against this; they argue that it is not the researcher's task to provide an index of transferability. Rather, they suggest, researchers should provide sufficiently rich data for the readers and users of research to determine whether transferability is possible. In this respect transferability requires 'thick description'.

Bogdan and Biklen (1992, p. 45) argue that, in qualitative research, we are more interested not with the issue of whether the findings are generalizable in the widest sense but with the question of the settings, people and situations to which they might be generalizable. Yin (2009) notes that qualitative research may be generalizable in terms of conforming to, or contributing to, a generalizable theory (see the discussion on case study at the end of this chapter). He also supports the use of replication studies here.

In naturalistic research, threats to external validity include (Lincoln and Guba, 1985, pp. 189, 300):

- selection effects (where constructs selected are only relevant to a certain group);
- setting effects (where the results are largely a function of their context);

- history effects (where the situations have been arrived at by unique circumstances and, therefore, are not comparable);
- construct effects (where the constructs used are peculiar to a certain group).

Onwuegbuzie and Leech (2006b, pp. 237–8) identify several threats to external validity in qualitative research that lie in the following fields:

1. *catalytic validity* (how far the research empowers the research community, or the effects of a piece of research);
2. *action validity* (how much use is made of the research findings by stakeholders and decision makers);
3. *investigation validity* (the ethical rigour, expertise, quality control and, indeed, personality of the researcher);
4. *interpretive validity* (how far the research catches the meanings and interpretations of the participants in the study);
5. *evaluative validity* (how far an evaluative structure (rather than a descriptive, interpretive or explanatory structure) can be applied to the research);
6. *consensual validity* (how far the 'competent others' agree on the interpretations made the research);
7. *population generalizability/ecological generalizability/temporal generalizability* (how successfully the researchers have kept within the bounds of generalizability/non-generalizability of their findings);
8. *researcher bias* (as for internal validity in qualitative research);
9. *reactivity* (as for internal validity in qualitative research);
10. *order bias* (where the order of the questions posed in an interview/observation/questionnaire affect the dependability of the results);
11. *effect size* (as for internal validity in qualitative research).

Researchers should decide, then, if they really seek generalizability and, if so, how to address this in the design of their research and the warrants brought forward for generalizability.

Construct validity

Construct validity is a fundamental type of validity. It is argued (Loevinger, 1957) that, in fact, construct validity is the queen of the types of validity because it subsumes other types of validity and because it concerns constructs or explanations rather than methodological factors, i.e. the meaning, definition and operationalization of factors.

A construct is an abstract which is theoretically derived; this separates it from other types of validity which deal in actualities – pre-defined content. In construct validity, agreement is sought on the 'operationalized' forms of a construct, clarifying what we mean when we work with this abstract construct, for example, is my understanding of this construct acceptable, fair in operationalizing the abstract construct, similar to that which is generally accepted to be the construct? For example, let us say that I wished to assess a child's intelligence (assuming, for the sake of this example, that it is a unitary quality). Intelligence is an abstract construct. I could say that I construe intelligence to be demonstrated in the ability to sharpen a pencil. How acceptable a construction and operationalization of, or an indicator of, intelligence is this? Is not intelligence something else (e.g. that which is demonstrated by a high score in an intelligence test)? To establish construct validity I would need to be assured that *my construction* of a particular issue is warranted, that proxies and indicators that I use for it in my research are warranted and agree with other constructions or theories of the same underlying abstract issue, for example, intelligence, creativity, anxiety, motivation.

Demonstrating construct validity means not only confirming the construction with that given in relevant literature or by the consistency of measures of the construct with other measures of that same construct; it also requires me to look for counter-examples which might falsify my construction. When I have balanced confirming and refuting evidence, I am in a position to demonstrate construct validity. I can stipulate what I take this construct to be. In the case of conflicting interpretations of a construct, I might have to acknowledge that conflict and then stipulate the interpretation that I shall use.

Addressing construct validity comprises two main stages:

Stage 1: Ensure that the construct has been correctly and adequately defined, including its key elements. This may require expert opinion, comparison with other tests of the construct in question, an exhaustive literature review and review of research in the field, a rooting in relevant theories of the construct in question.
Stage 2: Operationalize the construct fairly, so that the data-collection instruments fairly cover the construct and only the construct, i.e. rule out the effects of other possible constructs, which can be addressed using discriminant validity (see below), to show that the construct in question is different from other, possibly

similar, constructs. This can also be addressed by comparing the instrument used for data collection with other instruments purporting to address the construct, and by conducting correlational analysis of data from the instrument in question with data from other, related instruments.

Construct validity in quantitative research

Campbell and Fiske (1959), Brock-Utne (1996) and Cooper and Schindler (2001) suggest that construct validity is addressed by convergent and discriminant techniques. *Convergent techniques* imply that different methods for researching the same construct should give a relatively high inter-correlation, whilst *discriminant techniques* suggest that using similar methods for researching different constructs should yield relatively low inter-correlations, i.e. that the construct in question is different from other potentially similar constructs. Discriminant validity can be yielded by factor analysis, which clusters together similar issues and separates them from others (see Chapter 43). We discuss discriminant validity below.

Construct validity in qualitative research

In qualitative/ethnographic research, construct validity must demonstrate that the categories which the researchers are using are meaningful *to the participants themselves* (Eisenhart and Howe, 1992, p. 648), i.e. that they reflect the way in which the participants actually experience and construe the situations in the research, that they see the situation through the actors' eyes.

Threats to construct validity

There are several threats to construct validity (cf. Shadish *et al*., 2002, pp. 73–81), for example:

- poor definition of the construct, leading to incorrect inferences being made in its operationalization;
- failure to include all the elements of a construct;
- failure to identify what is and is not included in the construct (the boundaries of the construct);
- poor operationalization of the construct and its indicators/proxies (e.g. an intelligence test on its own is a highly selective construction of intelligence);
- confounding constructs: failure to address the fact that different constructs may be at work when one construct is being operationalized;
- failure to control out different factors (e.g. an intervention in a school to improve students' mathematics performance may find an improvement in mathematics scores, but this might overlook the fact that many students were taking private lessons in mathematics outside the school; see also Chapter 6 on causation);
- failure to separate one construct from another;
- false assumption that a construct can be measured by a single instrument (mono-method bias);
- failure to recognize that treatment may change the structure of a measure being used;
- failure to take account of participant reactivity to a situation, its novelty and processes.

Researchers have to be vigilant to ensure that these threats are addressed adequately.

Content validity

To demonstrate content validity, the instrument must show that it fairly and comprehensively covers the domain or items that it purports to cover (Carmines and Zeller, 1979, p. 20). It is unlikely that each issue will be able to be addressed in its entirety simply because of the time available or, for example, respondents' motivation to complete a long questionnaire, hence the researcher must ensure that the elements of the main issue to be covered in the research are both a fair representation of the wider issue under investigation (and its weighting) and that the elements chosen for the research sample are themselves addressed in depth and breadth. Careful sampling of items is required to ensure their representativeness.

For example, if the researcher wished to see how well a group of students could spell 1,000 words in French but decided to have a sample of only fifty words for the spelling test, then that test would have to ensure that the fifty words chosen fairly represented the range of spellings in the 1,000 words – maybe by ensuring that the spelling rules had all been included or that possible spelling errors had been covered in the test, in the proportions in which they occurred in the 1,000 words. The researcher would ensure that the population (the 1,000 words) covered all the aspects of spelling in which she was interested. Then she would randomly sample from the 1,000 items and then check that her fifty items selected fairly covered the 1,000 items.

The challenge here is to identify those characteristics required in the population (however defined: e.g. people, spelling items), i.e. to define the universe of content from which the sample will be drawn. In this respect expert opinion (jury validity) might be useful.

Convergent and discriminant validity

Convergent and discriminant validity are two sides of the same coin, and are both facets of construct validity. Convergent validity is demonstrated when two related or similar factors or elements of a particular construct

are shown (e.g. by measures or indicators) to be related or similar to each other, i.e. the results converge or are consistent with each other. Convergent validity is demonstrated when factors that *should* be related to each other are found, by indicators, actually to be related. Measures of correlation, regression, or factor analysis, are often used in quantitative research to demonstrate convergent validity. In qualitative research, where convergent validity is required to be shown, the researcher (e.g. using NVivo analysis and 'proximity searches', see Chapter 34) can show, by collating and collecting together data from people, groups, samples and subsamples, whether convergence has been found.

By contrast, discriminant (divergent) validity requires two or more unrelated items, attributes, elements or factors to be shown (e.g. by measurement) to be unrelated to, or different from, each other, i.e. difference is found where it should be found, even if those items at first seem to be similar. In quantitative research, statistics such as difference-testing (e.g. t-tests, chi-square tests, analysis of variance) are calculated. In qualitative research where discriminant validity is required, the researcher can examine negative cases, deviant cases and compare data from sub-groups of people, samples and sub-samples, cases and factors, to determine if, indeed, differences are found in terms of key factors, constructs, sub-elements or issues.

Convergent and discriminant validity can be addressed by mixed methods research. Here one can examine whether a set of data from one method accords with the data found by another method which focused on the same issues, variables or constructs. For example, the researcher could investigate whether the findings on, say, social class uptake of higher education in terms of cost–benefit to working-class students yield similar results from both qualitative and quantitative data. If they do, and if this was either predicted or supported by the literature, then one could suggest that convergent validity has been demonstrated. By contrast, let us say that the researcher hypothesized that family income and upward mobility aspirations for working-class students were not significantly related (the former being an index of wealth and the latter being an index of culture), and the data found two different, discordant results, then discriminant validity has been shown.

Convergent and discriminant validity draw on triangulation of methods, instruments, samples and theories. These important features of test construction are addressed in Chapter 27.

Criterion-related validity

Criterion-related validity concerns the detection of the presence or absence of suitable criteria that represent the construct in question, i.e. the appropriacy and suitability of the proxy or indicator being used. This can be addressed, for example, by administering the data-collection instrument (e.g. a test) to one group that is known to possess the construct in question, for example, extraversion, with such knowledge deriving from, say, experts or other data, and then looking to see which answers to which items in the test *did* correspond to the construct in question and which *did not*, in those participants known to possess the construct. Those items which have low correspondence are weeded out, leaving only those items which do correspond.

Criterion validity relates the results of one particular instrument to another external criterion. Within this type of validity there are two principal forms: predictive validity and concurrent validity. Predictive validity is achieved if the data acquired at the first round of research correlate highly with data acquired at a future date. For example, if the results of examinations taken by sixteen-year-olds correlate highly with the examination results gained by the same students when aged eighteen, then we might wish to say that the first examination demonstrated strong predictive validity.

In concurrent validity the data gathered from using one instrument must correlate highly with data gathered from using another instrument. For example, suppose I wished to research a student's problem-solving ability. I might observe the student working on a problem, or I might talk to the student about how she is tackling the problem, or I might ask the student to write down how she tackled the problem. Here I have three different data-collection instruments – observation, interview and documentation respectively. If the results all agreed – concurred – that, according to given criteria for problem-solving ability, the student demonstrated a good ability to solve a problem, I would be able to say with greater confidence (validity) that the student was good at problem solving than if I had arrived at that judgement simply from using one instrument.

Concurrent validity is similar to its partner – predictive validity – in its core concept (i.e. agreement with a second measure); what differentiates concurrent and predictive validity is the absence of a time element in the former; concurrence can be demonstrated simultaneously with another instrument.

An important partner to concurrent validity, which is also a bridge into later discussions of reliability, is triangulation, discussed later in this chapter.

Catalytic validity

Catalytic validity embraces the paradigm of critical theory discussed in Chapter 3 and the discussions of partisan research in that chapter. Put neutrally, catalytic

validity simply strives to ensure that research leads to action, echoing the paradigm of participatory research in Chapter 3. However, the story does not end there, for discussions of catalytic validity are substantive; like critical theory, catalytic validity often suggests an agenda. Lincoln and Guba (1986) suggest here that, in pursuing 'fairness', research should augment and improve participants' experience of the world, and should improve their empowerment. Lather (1986, 1991) and Kincheloe and McLaren (1994) suggest that the agenda for catalytic validity is to help participants understand their worlds in order to transform them, to bring about social justice, equality and empowerment. Catalytic validity, then, is intended to act as a spur to social change and transformation; its agenda is explicitly political, and it suggests the need to expose whose definitions of the situation are operating in the situation.

Catalytic validity is a major feature in critical theory, feminist research, critical race theory etc. (see Chapter 3), and, in these, it requires solidarity in the participants, an ability of the research to promote emancipation, autonomy and freedom within a just, egalitarian and democratic society (Masschelein, 1991), to reveal the distortions, ideological deformations and limitations that reside in research, communication and social structures (see also LeCompte and Preissle, 1993). Validity, it is argued (Mishler, 1990; Scheurich, 1996), is no longer an ahistorical given, but contestable, with definitions of valid research residing in the academic communities of the powerful. Lather (1986) calls for research to be emancipatory and to empower those who are being researched, suggesting that catalytic validity, akin to Freire's notion of 'conscientization', should empower participants to understand and transform their oppressed situation (discussed in Chapter 3 and its discussions of partisan research).

How defensible it is to suggest that researchers should have such ideological intents is a moot point; not to address this area is to perpetuate inequality by omission and neglect. Catalytic validity reasserts the centrality of ethics in the research process, as it requires researchers to interrogate their allegiances, responsibilities and self-interests (Burgess, 1989). We discuss this fully in Chapter 3.

Consequential validity

Partially related to catalytic validity is consequential validity, which argues that the ways in which research data are used (the consequences of the research) must be in keeping with the capability or intentions of the research, i.e. the consequences of the research do not exceed the capability of the research, and the action-related consequences of the research are both legitimate and fulfilled. Clearly, once the research is in the public domain the researcher has little or no control over how it is used. However, and this is often a political matter, research should not be used in ways in which it was not intended to be used, for example by exceeding the capability of the research data to make claims, by acting on the research in ways that the research does not support (e.g. by using the research for illegitimate epistemic support), by making illegitimate claims by using the research in unacceptable ways (e.g. by selection, distortion), and by not acting on the research in ways that were agreed, i.e. errors of omission and commission.

Cross-cultural validity

A considerable body of educational research seeks to understand the extent to which there are similarities and differences between cultures and their members. Matsumoto and Yoo (2006) identify four main phases of cross-cultural research:

- The first phase of making comparatively coarse cross-cultural comparisons of similarities and differences between cultures, though there is no attempt to demonstrate empirically (a) that differences found between groups are the result of cultural factors (pp. 234–5), and (b) what are the elements of the culture that have given rise to the differences.
- The second phase of 'identifying meaningful dimensions of cultural variability' (p. 235) identifies important dimensions of culture, and tests across cultures for the applicability, universality, extent and strength of these. An example of this are Hofstede's (1980) well-known dimensions of individualism–collectivism (see also Triandis, 1994), power–distance, uncertainty avoidance, masculinity–femininity and, later, long-term to short-term orientation (Hofstede and Bond, 1984). These studies have been criticized (Matsumoto and Yoo, 2006) for the assumption that: (a) countries are the same as cultures; (b) individual behaviour is the same as group behaviour (the ecological fallacy, discussed later); (c) there is a single or main culture in a country (i.e. overlooking differences within countries as well as between countries); and (d) attributing the causes of differences found between cultures to cultural sources rather than to other factors (e.g. economic factors, psychological factors).
- The third phase of cultural studies, in which theoretical models of culture and their influence on individuals are used to explain differences found between cultures, for example, Markus and Kitayama (1991)

on cognition, emotion and motivation, Nisbett (2005) on thought processes and cognition. This phase has been criticized for the limited empirical testing of 'cultural ingredients' (Matsumoto and Yoo, 2006).

- The fourth phase of establishing 'linkages' between empirical research on cultural variables and the models that hypothesize such linkages (Matsumoto and Yoo, 2006, p. 236).

For cross-cultural research to demonstrate validity, it is important to ensure that appropriate models of cross-cultural features and phenomena are developed, making clear their causal rootedness in cultural variables (rather than, e.g., psychological, economic or personality variables), that these models are operationalized into specific variables that constitute elements of culture, and that these are then tested empirically.

A major question to be faced by the cross-cultural researcher is the extent to which an instrument which has been developed, tested and validated in one country can be used in another culture or country. Are there sufficient similarities between the cultures or cultural properties (e.g. cultural 'universals') to enable the same instrument to be applied meaningfully in the other culture, given the particularities, uniqueness and sensitivities of each culture (e.g. Hilton and Skrutkowski, 2002; Sumathipala and Murray, 2006).

In conducting cross-cultural research, another fundamental issue to be addressed is in whose terms, constructs and definitions the researcher is working. This rehearses the 'emic'/'etic' discussion in Chapter 15, i.e. does the researcher use objective constructs, definitions, variables and elements of culture ('etic' views), or those that arise from the participants themselves ('emic views') (Hammersley, 2006, p. 6, 2013). Whose 'definition of the situation' drives the research? Are participants sufficiently aware of their own culture to be able to articulate it or, if the researcher uses/imposes her or his own construction of culture, is this a form of 'symbolic violence' to participants (Hammersley, 2006, p. 6)? In practice, the researcher can conduct pilot research (e.g. ethnographic research) to establish the categories, items and variables that are relevant, important and meaningful to participants, and then convert these into measurement scales for further investigation.

'Emic' research may be essential in cross-cultural research, as it is the locals who know more about their environment than an outside researcher (cf. Brock-Utne, 1996, p. 607) and who may know which are the important questions to ask in any environment; indeed she argues for the researcher being a local person rather than an outsider, as a local researcher will have more experience of, and hence more insight into, the local culture, though, of course, this should not blind the local researcher to the situation (p. 610). She gives a fascinating example of the interpretation of riddles in an African society; the outsider expatriate interprets them as entertainment and amusement, whereas the locals saw them as essential teaching and educational tools and promoters of cognitive development (pp. 610–12).

Items that are present in one culture may not be present in another, or may have different relevance, meanings or importance (Banville *et al*., 2000, p. 374). Banville *et al*. (2000) suggest the use of a team of experts in both cultures to work in parallel in order to establish the 'etic' constructs, and then they formulate questions for study that are subsequently operationalized into 'emic' constructs for each culture. This, they aver, avoids the danger of imposing an 'emic' culture from one culture as an 'etic' construct on another culture (p. 375) (see also Aldridge and Fraser, 2000, p. 127). Essentially the authors are arguing for ensuring the relevance of the instrument for all the target cultures, by including 'emic' and 'etic' elements.

It is important to address meaningfulness and relevance in cross-cultural research: whilst a construct or element of culture may be found in two cultures, it may have different meanings, weight or significance in the two cultures, i.e. the *presence* alone of a factor may not be sufficient in cross-cultural research.

Threats to validity in cross-cultural research may lie in many areas, for example:

- failure to operationalize elements of cultures into researchable variables;
- problems of whose construction of 'culture' to adopt: 'emic' and/or 'etic' research;
- false attribution of causality for differences found between groups to cultural factors rather than non-cultural factors, for example, economic factors, affluence, demography, biological features of people, climate, personality, religion, educational practices, personal/subjective perceptions of the research, contextual but non-cultural variables (Alexander, 2000; Matsumoto and Yoo, 2006);
- the ecological fallacy: the error of the ecological fallacy is made where

  relationships that are found between aggregated data (e.g. mean scores) are assumed to apply to individuals, i.e. one infers an individual or particular characteristic from a generalization. It assumes that the individuals in a group exhibit the same features of the whole group taken together (a form of stereotyping).
  (Morrison, 2009, p. 62)

The caution here is to avoid assuming that what one finds at a group level is necessarily the same as that which one would find at an individual level;

- the directions of causality, for example, whether culture influences individual behaviour or vice versa, or both;
- sampling, for example, much cross-cultural research involves using groups of university students, or – as in the case of Hofstede (1980) – individual companies, and it is dangerous to generalize more widely from these. Further, some studies do not have samples that are matched in terms of size or characteristics of the sample;
- instrument problems: different groups may not understand, or have different understandings of, the language/issues/instruments used for gathering data;
- problems of convergent validity (where several items that are supposed to be measuring the same construct or variable do not yield strong inter-correlations);
- problems of discriminant validity (where items that are supposed to be measuring different constructs or variables yield strong inter-correlations);
- problems of equivalence (where the same meaning and significance is not given to concepts, constructs, language, sampling, methods in different cultures, such that meaningful comparisons cannot be made between cultures);
- problems of conceptual equivalence (where items are unrelated or relatively unimportant or meaningless to one or more groups) (e.g. Aldridge and Fraser, 2000, p. 111);
- problems of psychological equivalence, where the psychological connotations or referents in the original language may be different from those in the translated language, giving rise to differences in results that are attributable to factors other than cultural (Liu, 2002; Riordan and Vandenburg, 1994);
- problems of meaning equivalence: using similar words in the two languages but which connote different interpretations or meanings;
- failure of the instruments to take account of different frames of reference of the different cultural groups (Riordan and Vandenburg, 1994);
- failure of groups to understand the measures, instruments, language, meaning or research, i.e. the same items may be interpreted differently by different groups;
- failure to accord equal significance to items (factors might be found to be present in different cultures, but some cultures accord those factors much more importance than others, e.g. in measures of personality such as the Big Five factors of personality) (Matsumoto and Yoo, 2006, p. 240);
- failure to accord equal relevance and meaning to the same construct or item in different cultures;
- measurement equivalence;
- linguistic equivalence (where translated versions of an instrument carry the same meaning as in the original, and which will be understood in the same way by members of different cultures);
- response bias, in which members of different cultures respond in systematically different ways to items, elements, constructs or scales in the instrument in ways that are meaningful to their own cultures, situations or contexts (Riordan and Vandenburg, 1994; Aldridge and Fraser, 2000, p. 127). For example: (a) some cultures may give more weight to socially desirable responses or to responses that make the participants look good (Liu, 2002, p. 82); (b) some cultures may give more weight to categories of 'agree' rather than 'disagree' in responses; (c) some cultures may consider it undesirable to use extreme ends of a measurement scale such as 'strongly agree' or 'strongly disagree', or indeed some cultures may deliberately value the use of extreme categories, such as those that emphasize status, masculinity and power (Matsumoto and Yoo, 2006);
- preparation of participants – giving advance organizers or suggestions to participants before administering an instrument ('priming') (Matsumoto and Yoo, 2006) – may give rise to different responses;
- problems with the researcher who may not speak the language(s) of the participants, or whose participants may be insufficiently articulate or literate to engage in respondent validation.

There are several techniques that researchers can use to address validity in cross-cultural research. For instruments such as questionnaires, a common practice is to use 'back-translation', undertaken by bilinguals or those with a sound ability in the second as well as the first language (cf. Brislin, 1970; Vallerand *et al*., 1992; Banville *et al*., 2000; Cardinal *et al*., 2003). Here the original version of the instrument (say, a questionnaire in English) is translated into the other language required (say, Chinese). Then the Chinese version is given to a third party who does not have sight of the original English version, and that third party translates the Chinese version back into English. The two English versions (the original and the resultant back-translation) are then compared to check whether the meanings (and, in a few cases, the exact language) are the same. If the meanings in the two English versions are the same (semantic equivalence) then the Chinese version is said to be acceptable; if the meanings in the two English

versions are discrepant then there may be a problem in the Chinese, and the Chinese translation is revisited to make changes to it.

Liu (2002) suggests that translators should be familiar with the subject matter, and, if possible, instrumentation. Banville *et al*. (2000) report the use of professional translators instead of simply back-translation, in order to ensure discriminability of similar items in translation, and they indicate that translation should precede the conduct of the empirical research and that translated instruments should be piloted to determine their suitability for the target population.

A variant of this, to ensure even greater validity and reliability of the translated version, is to have more than one person doing the translation into the new language (each person is unknown to the other) and similarly for the back-translation into the original language, as this avoids possible bias in having only a single translator at each stage (Banville *et al*., 2000, p. 379). In this instance, the two translators at each stage should compare their translations and discuss any differences found in meaning or language.

Aldridge and Fraser (2000) note that there may be no equivalent words in the target translated language, and this may mean that there have to be rewordings of the original language in order to reach a compromise statement in the instrument (e.g. a questionnaire) that fits both languages. For example, in translating the English phrase 'how much' into Chinese, the Chinese characters change, depending on the topic in hand. Whilst back-translation keeps the original language as the language of reference, in fact compromises may have to be made in both the original and the translated language, in order to ensure commonality or equivalence of meaning, i.e. the original and the translated language are equally important and must be user-friendly to all groups (Liu, 2002, p. 81). Liu also suggests that it is useful to keep the original language in active rather than passive voice, simple and short sentences, avoiding colloquialisms, idioms and using specific terms and familiar rather than abstruse words (see also Hilton and Skrutkowski, 2002).

Banville *et al*. (2000) provide a useful seven-step approach from Vallerand (1989) to translating and using instruments in cross-cultural research:

Step 1: Prepare a preliminary version of the instrument using the back-translation technique.

Step 2: Evaluate the preliminary versions (to check that the back-translated version is acceptable, or to adjudicate between different versions of the back-translated items) and prepare an experimental version of the instrument using a committee of experts (3–5 persons) to conduct such a review, thereby avoiding possible bias by a single researcher (see also Vallerand *et al*., 1992; Liu, 2002, p. 82).

Step 3: Pre-test the experimental version using a random survey approach, to check the clarity of the instructions and the appropriateness of the instrument.

Step 4: Evaluate the content and concurrent validity of the instrument using bilingual participants to check whether they are answering both versions in the same way, and to check the appropriateness of the instrument (using between twenty and thirty participants). Participants answer both versions of the instrument (i.e. both languages). Content validity can be assessed qualitatively (expert review) and concurrent validity can be assessed quantitatively (e.g. by difference testing or correlational analysis).

Step 5: Conduct a reliability analysis to check for internal validity and stability over time (looking for high reliability coefficients: Cronbach alphas and correlations respectively), and to check the suitability of the instrument. Remove items with low reliability.

Step 6: Evaluate the construct validity of the instruments (through factor analysis, inter-scale correlations and to test the hypothesis that stems from theory).

Step 7: Establish norms of the scales/measures by selecting the population from which the sample will be drawn, by statistical indices, and by calculating means, standard deviations and standardized (z) scores, used with a large number of people in order to establish the stability of the norms (see Chapters 40–43 of the present volume).

Step 4 uses bilingual participants to undertake both versions (both languages), so that their two sets of answers can be compared for discrepancies (see also Liu, 2002, pp. 81–2). This may not be feasible for sole researchers, who may not have access to a sufficiently large group of bilingual participants, but only to people who can translate rather than who are fully bilingual and expert in both cultures. (For an example of the use of this technique, see Cothran *et al*., 2005.)

In order to avoid bias in cross-cultural research, the researcher can also use a multi-instrument approach with different-sized samples for different instruments (Aldridge and Fraser, 2000; Aldridge *et al*., 1999; Sumathipala and Murray, 2006). A multi-method approach provides triangulation and concurrent validity

and gives a closer, more authentic meaning to the phenomenon or culture (particularly when qualitative data combine with quantitative data).

Qualitatively speaking, the researcher has to ensure that: (a) the meanings, definitions and constructs which are being used are understood similarly by the members of the different cultures being investigated (the equivalence issue); (b) these are given sufficient relevance, meaningfulness and weight in the different cultures for them to be suitable for investigation (or, indeed, the research may be intended to discover the relevance, meaningfulness and weight of these in the different cultures); (c) the research includes items that are meaningful, relevant and significant to participants; and (d) the research draws on both 'emic' and 'etic' analysis and constructs as appropriate.

Quantitatively speaking, there are several ways in which the cross-cultural validity of measures can be addressed. We discuss these below. Essentially the purpose is to test the instrument on the different cultures to see if the reliability, items, clusters of items into factors and suitability of the items are acceptable in both cultures; an instrument that is suitable, reliable and valid in one culture may not be in another (Cothran *et al*., 2005, p. 194).

Factor analysis enables the researcher to examine the factor structure of the instrument. A suitable instrument for cross-cultural research should ensure that: (a) the same *factors* are extracted from the same instrument with the different groups of participants; (b) the same *variables* are included in these factors with the different groups of participants; (c) the same *loadings* (e.g. weightings) of each variable are loaded onto each factor (see Chapter 43). One has to exercise discretion here, as, clearly, the results will not be identical for each group of participants. However, if there are gross discrepancies found between factors, variables included, and loadings of each variable, then the researcher will need to consider whether the instrument is sufficiently valid, or whether some items will need to be excluded or replaced.

Inter-correlations of variables (alphas) (discussed below in section on 'Reliability') can be conducted to see whether: (a) the item-to-whole reliability correlation coefficient is the same for the different groups of participants; (b) the overall reliability level (the alpha) is sufficiently high for items to be included (see Chapter 40). A suitable instrument will ensure that the coefficient of correlation for each item to the whole is sufficiently high (e.g. ≥ 0.67), or the overall alphas for the sections of the instrument are sufficiently high (e.g. ≥ 0.67) to be retained. Items with low correlations should be considered for removal. Hence the researcher will need to test his/her instrument in the groups concerned (e.g. groups of members of different cultures) in order to conduct such pilot testing. In this case it is advisable to include no fewer than thirty people in each of the pilot groups.

Items which, the researcher hypothesizes, should be strongly correlated, i.e. convergent validity: measuring the same construct, factor or trait (Rohner and Katz, 1970, p. 1069), should have high correlation coefficients. Items which, the researcher hypothesizes, should have very low correlation coefficients, i.e. discriminant validity: measuring unrelated constructs, factors or traits (p. 1069), should have low correlation coefficients. Alternatively, instead of using correlations, the researcher can conduct difference testing (e.g. t-tests, ANOVA see Chapter 41) to discover: (a) whether items which, he/she hypothesizes, should be similar to each other (convergent validity), in reality show no statistically significant difference or very small effect size; and (b) whether items which, he/she hypothesizes, should be different from each other (discriminant validity), in reality are statistically significantly different from each other or have high effect sizes (see Keet *et al*. (1997) for an example of using correlational analysis, t-tests and factor analysis to establish validity in cross-cultural research).

Watkins (2007, pp. 305–6) suggests that meta-analysis can be used to examine the cross-cultural relevance of variables to the participating groups. This is a statistical procedure in which the researcher selects and combines empirical studies that satisfy criteria for inclusion in respect of the hypotheses under investigation (e.g. they are quantitative, include relevant variables, include scales and measures that can be combined from different studies, include identified samples and include correlational analysis of items). Then the researcher calculates average correlations and effect sizes from the studies (bearing in mind the likely different sample sizes), and then judges whether the correlations and effect sizes found are sufficiently strong for items to be retained in the researcher's own research (on how to conduct a meta-analysis, see Glass *et al*., 1981; Hattie, 2009; Cumming, 2012).

Cross-cultural validity, like other forms of research, should be cautious in making generalizations from small samples, in avoiding claims about whole cultures or countries from limited or selective samples and in imposing instruments from one culture on another – however well they might be translated. Matsumoto and Yoo (2006) suggest that cross-cultural data are 'nested' (p. 246), i.e. there are data at several levels: individual, group, cultures, societies, ecologies. This points us to the statistical technique of multilevel modelling.

Cultural validity

Related to cross-cultural research and ecological validity (see below) is cultural validity (Morgan, 1999). This is particularly an issue in cross-cultural, intercultural and comparative kinds of research, where the intention is to shape research so that it is appropriate to the culture of the researched, and where the researcher and the researched are members of different cultures. Cultural validity is defined as 'the degree to which a study is appropriate to the cultural setting where research is to be carried out' (Joy, 2003, p. 1; see also Stuchbury and Fox, 2009, p. 494). Cultural validity, Morgan (1999) suggests, applies at all stages of the research, and affects its planning, implementation and dissemination. It involves a degree of sensitivity to the participants, cultures and circumstances being studied. Morgan (2005) writes that:

> cultural validity entails an appreciation of the cultural values of those being researched. This could include: understanding possibly different target culture attitudes to research; identifying and understanding salient terms as used in the target culture; reviewing appropriate target language literature; choosing research instruments that are acceptable to the target participants; checking interpretations and translations of data with native speakers; and being aware of one's own cultural filters as a researcher.
>
> (Morgan, 2005, p. 1)

Joy (2003, p. 1) presents twelve important questions that researchers in different cultural contexts may face, to ensure that research is culture-fair and culturally sensitive:

1. Is the research question understandable and of importance to the target group?
2. Is the researcher the appropriate person to conduct the research?
3. Are the sources of the theories that the research is based on appropriate for the target culture?
4. How do researchers in the target culture deal with the issues related to the research question (including their method and findings)?
5. Are appropriate gatekeepers and informants chosen?
6. Are the research design and research instruments ethical and appropriate according to the standards of the target culture?
7. How do members of the target culture define the salient terms of the research?
8. Are documents and other information translated in a culturally appropriate way?
9. Are the possible results of the research of potential value and benefit to the target culture?
10. Does interpretation of the results include the opinions and views of members of the target culture?
11. Are the results made available to members of the target culture for review and comment?
12. Does the researcher accurately and fairly communicate the results in their cultural context to people who are not members of the target culture?

Ecological validity

In education, ecological validity is particularly important and useful in charting how policies are actually happening 'at the chalk face' (Brock-Utne, 1996, p. 617). It concerns examining and addressing the specific characteristics of a particular situation, for example, how policies are actually impacting in practice (p. 617) rather than simply assuming that policies are implemented in the ways intended or in the ways that the powerful groups intended (those at 'the top of the hierarchy of credibility'; p. 618).

Ecological validity requires the specific factors of research sites – schools, universities, regions etc. – to be included and taken into account in the research. In this respect it is more sympathetic to qualitative research and 'thick description' (Geertz, 1973) than those forms of quantitative research variables which seek to isolate, control out and manipulate variables in contrived settings. The ethical tension is raised in ecological validity between the need to provide rich descriptions of characteristics of a situation or institution and the increased likelihood that this will lead to the situation or institution being able to be identified and anonymity breached (Brock-Utne, 1996, p. 618).

To demonstrate ecological validity, it is important to include and address in the research as many as possible of the characteristics and factors of a given situation. The intention is to give accurate portrayals of the realities of social situations in their own terms, in their natural or conventional settings. The difficulty with this is that the more characteristics are included and described, the harder it is to abide by central ethical tenets of much research – non-traceability, anonymity and non-identifiability.

Ecological validity raises the issues of external validity: the extent to which characteristics of one situation or behaviour observed in one setting can be transferred or generalized to another situation; how far fidelity to one specific set of circumstances can apply to others.

14.6 Triangulation

In its original and literal sense, triangulation is a technique of physical measurement: maritime navigators, military strategists and surveyors, for example, use (or used to use) several locational markers in their endeavours to pinpoint a single spot or objective. By analogy, triangular techniques in the social sciences attempt to map out, or explain more fully, the richness and complexity of human behaviour by studying it from more than one standpoint and, in so doing, by making use of both quantitative and qualitative data. Triangulation is a powerful way of demonstrating concurrent validity.

For example, the advantages of the mixed methods approach in social research are manifold and we examine two of them. First, it has been observed that as research methods act as filters through which the environment is selectively experienced, they are never atheoretical or neutral in representing the world of experience (see Chapter 1). Exclusive reliance on one method, therefore, may bias or distort the researcher's picture of the particular slice of reality she is investigating. She needs to be confident that the data generated are not simply artefacts of one specific method of collection (Lin, 1976). Such confidence can be achieved, as far as nomothetic research is concerned, when different methods of data collection yield substantially the same results. (Where triangulation is used in interpretive research to investigate different actors' viewpoints, the same method, e.g. accounts, will naturally produce different sets of data.)

Second, the more the methods contrast with each other, the greater is the researcher's confidence. If, for example, the outcomes of a questionnaire survey correspond to those of an observational study of the same phenomenon, the more the researcher can be confident about the findings. Or, more extremely, where the results of a rigorous experimental investigation are replicated in, say, a role-playing exercise, the researcher will experience even greater assurance. If findings are artefacts of method, then the use of contrasting methods considerably reduces the chances of any consistent findings being attributable to similarities of method (Lin, 1976). The use of triangular techniques, it is argued, can help to overcome the problem of 'method-boundedness'; indeed Chapter 2 demonstrates the value of combining qualitative and quantitative methods. In its use of mixed methods, triangulation may utilize either normative or interpretive techniques, or it may draw on methods from both these approaches and use them in combination.

Types of triangulation and their characteristics

Triangulation is often characterized by a mixed methods approach to a problem in contrast to a single-method approach. Denzin (1970) has, however, extended this view of triangulation to take in several other types as well as the mixed methods kind which he terms 'methodological triangulation', including:

- time triangulation: this takes into consideration the factors of change and process by utilizing cross-sectional and longitudinal designs. Kirk and Miller (1986) suggest that diachronic reliability seeks stability of observations over time, whilst synchronic reliability seeks similarity of data gathered in the same time;
- space triangulation: this attempts to overcome the parochialism of studies conducted in the same country or within the same subculture by making use of cross-cultural techniques;
- combined levels of triangulation: this uses more than one level of analysis from the three principal levels used in the social sciences, namely, the individual level, the interactive level (groups) and the level of collectivities (organizational, communitarian, cultural or societal);
- theoretical triangulation: this draws upon alternative or competing theories in preference to utilizing one viewpoint only;
- investigator triangulation: this engages more than one observer, and data are discovered independently by more than one observer (Silverman, 1993, p. 99);
- methodological triangulation: this uses either (a) the same methodology on different occasions or (b) different methods on the same object of study.

We can add to these:

- paradigm triangulation: different paradigms used in the same study;
- instrument triangulation: data-collection instruments;
- sampling triangulation: different samples and subsamples.

Many studies in the social sciences are conducted at one point only in time, thereby excluding effects of social change and process. Time triangulation goes some way to rectifying these omissions by making use of longitudinal approaches. Longitudinal studies collect data from the same group at different points in time. The use of panel studies and trend studies also address the time dimension (see Chapter 17). The former

compare the same measurements for the same individuals in a sample at several different points in time, and the latter examine selected processes continually over time. The weaknesses of each of these methods can be strengthened by using a combined approach to a given problem.

Space triangulation attempts to overcome the limitations of studies conducted within one culture or subculture (cf. Smith, 1975), as behavioural sciences are culture-bound and subculture-bound rather than being automatically true of any societies. Cross-cultural studies may involve testing theories among different people, as in Piagetian psychology, or they may measure differences between populations by using several different measuring instruments. We have addressed cultural validity earlier.

Social scientists are concerned with the individual, the group and society. These reflect three levels of analysis adopted by researchers. Those who are critical of research argue that some of it uses the wrong level of analysis, for example individual when it should be societal, or that it limits itself to one level only when a more meaningful picture would emerge by using more than one level. Smith (1975) extends this analysis and identifies seven possible levels: the aggregative or individual level, and six levels which characterize the collective as a whole, and do not derive from an accumulation of individual characteristics. The six are:

- group analysis (the interaction patterns of individuals and groups);
- organizational units of analysis (units which have qualities not possessed by the individuals making them up);
- institutional analysis (relationships within and across the legal, political, economic and familial institutions of society);
- ecological analysis (concerned with spatial explanation);
- cultural analysis (concerned with the norms, values, practices, traditions and ideologies of a culture); and
- societal analysis (concerned with gross factors such as urbanization, industrialization, education, wealth, etc.).

Studies combining several levels of analysis are useful.

Theoretical triangulation requires researchers to look at a phenomenon through different theoretical lenses. Researchers are sometimes taken to task for their rigid adherence to one particular theory or theoretical orientation to the exclusion of competing theories. Indeed a major function of research is to test competing theories.

Investigator triangulation refers to the use of more than one observer (or participant) in a research setting. Observers working on their own each have their own observational styles and this is reflected in the resulting data. The careful use of two or more observers or participants independently can lead to more valid and reliable data, checking divergences between researchers and leading to minimal divergence, i.e. reliability.

Denzin (1970) identifies two categories in methodological triangulation: 'within methods' triangulation and 'between methods' triangulation. Triangulation within methods concerns the replication of a study as a check on reliability and theory confirmation. Triangulation between methods involves the use of more than one method in the research. As a check on validity, the 'between methods' approach embraces the notion of convergence between independent measures of the same objective (Campbell and Fiske, 1959). Triangulation bridges issues of reliability and validity.

Triangular techniques are suitable when a more holistic view of educational outcomes is sought, or where a complex phenomenon requires elucidation. Triangulation is useful when an established approach yields a limited and frequently distorted picture. It can also be useful where a researcher is engaged in case study, a particular example of complex phenomena (Adelman *et al*., 1980).

Triangulation is not without its critics. For example, Silverman (1985) suggests that the very notion of triangulation is positivistic, and that this is exposed most clearly in *data triangulation*, as it suggests that a multiple data source (concurrent validity) is superior to a single data source or instrument. The assumption that a single unit can always be measured more than once violates the interactionist principles of emergence, fluidity, uniqueness and specificity (Denzin, 1997, p. 320). Further, Patton (1980) suggests that even having multiple data sources, particularly of qualitative data, does not ensure consistency or replication. Fielding and Fielding (1986) hold that methodological triangulation does not necessarily increase validity, reduce bias or bring objectivity to research. Further, triangulation suggests that there is only one correct final position, conclusion or focus (Tracy, 2010); in qualitative research this may not be the case.

With regard to investigator triangulation, Lincoln and Guba (1985, p. 307) contend that it is erroneous to assume that one investigator will corroborate another, nor is this defensible, particularly in qualitative, reflexive inquiry. They extend their concern to include theory and methodological triangulation, arguing that the search for theory and methodological triangulation is

epistemologically incoherent and empirically empty (see also Patton, 1980). No two theories, it is argued, will ever yield a sufficiently complete explanation of the phenomenon being researched.

These criticisms are trenchant, but they have been answered equally trenchantly by Denzin (1997). In naturalistic inquiry, Lincoln and Guba (1985, p. 315) suggest that triangulation is intended as a check on data, whilst member checking, an element of credibility, can be used as a check on members' constructions of data.

14.7 Ensuring validity

It is easy to slip into invalidity; it can enter at every stage of a piece of research. The attempt to build out invalidity is essential if the researcher is to have confidence in the elements of the research plan, data acquisition, data-processing analysis, interpretation and its ensuing judgement.

At the *design stage*, threats to validity can be minimized by:

- choosing an appropriate timescale;
- ensuring that there are adequate resources for the required research to be undertaken;
- selecting an appropriate methodology for investigating and answering the research questions;
- selecting appropriate instrumentation for gathering the type of data required;
- using an appropriate sample (e.g. which is representative, not too small nor too large);
- demonstrating internal, external, content, concurrent and construct validity; 'operationalizing' the constructs fairly;
- ensuring reliability in terms of stability (consistency, equivalence, split-half analysis of test material);
- selecting appropriate foci to answer the research questions;
- devising and using appropriate instruments (e.g. to catch accurate, representative, relevant and comprehensive data; ensuring that readability levels are appropriate; avoiding any ambiguity of instructions, terms and questions; using instruments that will catch the complexity of issues; avoiding leading questions; ensuring that the level of test is appropriate – neither too easy nor too difficult; avoiding test items with little discriminability; avoiding making the instruments too short or too long; avoiding too many or too few items for each issue);
- avoiding a biased choice of researcher or research team (e.g. insiders or outsiders as researchers).

At the *data-gathering stage*, threats to validity can be minimized by:

- reducing the Hawthorne effect (see the accompanying website);
- minimizing reactivity effects (respondents behaving differently when subjected to scrutiny or being placed in new situations, e.g. the interview situation – we distort people's lives in the way we go about studying them (Lave and Kvale, 1995, p. 226));
- trying to avoid dropout rates among respondents;
- taking steps to avoid non-return of questionnaires;
- avoiding having too long or too short an interval between pre-tests and post-tests;
- ensuring inter-rater reliability;
- matching control and experimental groups fairly;
- ensuring standardized procedures for gathering data or for administering tests;
- building on the motivations of the respondents;
- tailoring the instruments to the concentration span of the respondents and addressing other situational factors (e.g. health, environment, noise, distraction, threat);
- addressing factors concerning the researcher (particularly in an interview situation), for example, the attitude, gender, ethnicity, age, personality, dress, comments, replies, questioning technique, behaviour, style and non-verbal communication of the researcher.

At the *data-analysis stage*, threats to validity can be minimized by:

- using respondent validation;
- avoiding subjective interpretation of data (e.g. being too generous or too ungenerous in the award of marks), i.e. lack of standardization and moderation of results;
- reducing the halo effect, where the researcher's knowledge of the person or knowledge of other data about the person or situation exerts an influence on subsequent judgements;
- using appropriate statistical treatments for the level of data (e.g. avoiding applying techniques from ratio scales data to ordinal data or using incorrect statistics for the type, size, complexity, sensitivity of data);
- recognizing spurious correlations and extraneous factors which may be affecting the data;
- avoiding poor coding of qualitative data;
- avoiding making inferences and generalizations beyond the capability of the data to support such statements;

- avoiding the equating of correlations and causes;
- avoiding selective use of data;
- avoiding unfair aggregation of data (particularly of frequency tables);
- avoiding unfair telescoping of data (degrading the data);
- avoiding Type I and/or Type II errors.

At the *data-reporting stage*, threats to validity can be minimized by:

- avoiding using data selectively and unrepresentatively (e.g. accentuating the positive and neglecting or ignoring the negative);
- indicating the context and parameters of the research in the data collection and treatment, the degree of confidence which can be placed in the results, the degree of context-freedom or context-boundedness of the data (i.e. the level to which the results can be generalized);
- presenting the data without misrepresenting its message;
- making claims which are sustainable by the data;
- avoiding inaccurate or wrong reporting of data (technical or orthographic errors);
- ensuring that the research questions are answered; releasing research results neither too soon nor too late.

Having identified where invalidity might obtain, the researcher can take steps to ensure that, as far as possible, it has been minimized in all areas of the research.

14.8 Reliability

Reliability is essentially an umbrella term for dependability, consistency and replicability over time, over instruments and over groups of respondents. Can we believe the results? Reliability is concerned with precision and accuracy: some features, for example, height, can be measured precisely, whilst others, for example, musical ability, cannot. For research to be reliable it must demonstrate that if it were to be carried out on a similar group of respondents in a similar context (however defined), then similar results would be found. Guba and Lincoln (1994) suggest that the concept of reliability is largely positivist. Whilst widely held views of reliability may seem to adhere to positivism rather than to qualitative research, it is not exclusively so; qualitative research must be as reliable as positivist and post-positivist research, though in different ways: the canons of reliability and the types of reliability differ in quantitative and qualitative research. Similarly, it is simply not the case that qualitative or quantitative research, per se, guarantees reliability or that it is an irrelevance in qualitative research (Brock-Utne, 1996, p. 613). Reliability is relevant to both quantitative and qualitative research.

14.9 Reliability in quantitative research

In quantitative research and qualitative research which seeks trends, patterns, predictability and control (e.g. Miles and Huberman, 1994), there are three principal types of reliability: stability, equivalence and internal consistency (Carmines and Zeller, 1979). Here reliability concerns the research situation (e.g. the context of, or the conditions for, a test), factors affecting the researcher or participants, and the instruments for data collection themselves.

Reliability as stability

Reliability as stability is a measure of consistency over time, over similar samples and over the uses of the instrument in question. A reliable instrument in a piece of research yields similar data from similar respondents over time. A leaking tap which leaks one litre each day is leaking reliably, whereas a tap which leaks one litre some days and two litres on another, is not. In the experimental and survey models of research this would mean that if a test and then a re-test were undertaken within an appropriate time span, with no changes having occurred, then similar results should be obtained. The researcher has to decide what is an appropriate length of time; too short a time and respondents may remember what they said or did in the first test situation; too long a time and there may be extraneous effects operating to distort the data (e.g. maturation in students, outside influences on the students). A researcher seeking to demonstrate this type of reliability will have to choose an appropriate timescale between the test and re-test. Correlation coefficients can be calculated for the reliability of pre- and post-tests, using formulae which are readily available in texts on statistics and test construction and on Internet sites.

In addition to stability over time, reliability as stability can also be stability over a similar sample. For example, we would assume that if we were to administer a test or a questionnaire simultaneously to two groups of students who were very closely matched on significant characteristics (e.g. age, gender, ability etc. – whatever characteristics are deemed to have a significant bearing on the responses), then similar results (on a test) or responses (to a questionnaire) would be obtained. The

correlation coefficient on this form of the test/re-test method can be calculated either for the whole test or for sections of the questionnaire (e.g. by using a correlation statistic or a t-test as appropriate). The correlation coefficient can be found and should be high for reliability to be guaranteed. This form of reliability over a sample is particularly useful in piloting tests and questionnaires.

In using the test/re-test method, care has to be taken to ensure the following (Cooper and Schindler, 2001, p. 216):

- the time period between the test and re-test is not so long that situational factors may change;
- the time period between the test and re-test is not so short that the participants will remember the first test or that intervention effects will be too strong to be reliable (e.g. the Hawthorne effect and the immediacy effect);
- the participants may have become interested in the field and may have followed it up themselves between the test and the re-test times.

Reliability as equivalence

There are two main kinds of reliability as equivalence. Reliability may be achieved, first, through using equivalent forms (also known as 'alternative forms') of a test or data-gathering instrument. If an equivalent form of the test or instrument is devised and yields similar results, then the instrument can be said to demonstrate this form of reliability. For example, the pre-test and post-test in an experiment are predicated on this type of reliability, being alternate forms of instrument to measure the same issues. This type of reliability might also be demonstrated if the equivalent forms (e.g. items) of a test or other instrument yield consistent results if applied simultaneously to matched samples (e.g. two random samples in a survey). Here reliability can be measured through a difference test (e.g. a t-test or a Mann–Whitney U test), through the demonstration of a high correlation coefficient, similar means and standard deviations between two groups.

Second, reliability as equivalence may be achieved through inter-rater reliability. If more than one researcher is taking part in a piece of research then, human judgement being fallible, agreement between all researchers must be achieved, through ensuring that each researcher enters data in the same way. This is particularly pertinent to a team of researchers gathering structured observational or semi-structured interview data where each member of the team must agree on which data to enter into which categories. For observational data, such reliability is addressed in training sessions for researchers, for example, working on video material to ensure parity in how to enter data.

At a simple level one can calculate the inter-rater agreement as a percentage:

$$\frac{\text{Number of actual agreements}}{\text{Number of possible agreements}} \times 100$$

Robson (2002, p. 341) sets out a more sophisticated way of measuring inter-rater reliability in coded observational data, and his method can be used with other types of data.

Reliability as internal consistency

Whereas the test/re-test method and the equivalent forms method of demonstrating reliability require the tests or instruments to be done twice, demonstrating internal consistency demands that the instrument or tests be run once only through the split-half method.

Let us imagine that a test is to be administered to a group of students. Here the test items are divided into two halves, ensuring that each half is matched in terms of item difficulty and content. Each half is marked separately. If the test demonstrates split-half reliability, then the marks obtained on each half should correlate highly with each other. Any student's marks on the one half should match his or her marks on the other half. This can be calculated using the Spearman-Brown formula:

$$\text{Reliability} = \frac{2r}{1+r}$$

where r = the actual correlation between the halves of the instrument.

This calculation requires a correlation coefficient to be calculated, for example, a Spearman rank order correlation or a Pearson product moment correlation (Chapter 40). Let us say that using the Spearman-Brown formula, the correlation coefficient is 0.85; in this case the formula for reliability is set out thus:

$$\text{Reliability} = \frac{2x0.85}{1+0.85} = \frac{1.70}{1.85} = 0.919$$

Given that the maximum value of the coefficient is 1.00, we can see that the reliability of this instrument, calculated using the split-half reliability testing, is very high.

This type of reliability assumes that the test can be split into two matched halves; many tests have a gradient of difficulty or different items of content in each half. If this is the case and, for example, the test contains twenty items, then the researcher, instead of splitting the test into two by assigning items 1–10 to one half and items 11–20 to the second half, may assign all the even-numbered items to one group and all the odd-numbered items to another. This moves to the two

halves being matched in terms of content and cumulative degrees of difficulty.

An alternative measure of reliability as internal consistency is the Cronbach alpha, frequently referred to simply as the alpha coefficient of reliability, or simply the alpha. The Cronbach alpha provides a coefficient of inter-item correlations, i.e. the correlation of each item with the sum of all the other relevant items. This is useful for multi-item scales and is a measure of the internal consistency among the *items* (not, for example, the people). We address the alpha coefficient and its calculation in Chapter 40.

Ary *et al*. (2002, pp. 262–3) suggest that reliability of a data-collection instrument is a function of:

- the length of the data-collection instrument (e.g. a test);
- the heterogeneity of the group being investigated (the greater the heterogeneity, the greater the reliability);
- the abilities of the participants;
- the methods of testing for reliability;
- the nature of the variable that is being measured or investigated.

Reliability, thus construed, makes several assumptions, for example, that instrumentation, data and findings should be controllable, predictable, consistent and replicable. This pre-supposes a particular style of research, for example, positivist or post-positivist. Cooper and Schindler (2001, p. 218) suggest that, here, reliability can be improved by: minimizing any external sources of variation – standardizing and controlling the conditions under which the data collection and measurement take place; training the researchers in order to ensure consistency (inter-rater reliability); widening the number of items on a particular topic; excluding extreme responses from the data analysis (e.g. outliers, which can be done with SPSS).

14.10 Reliability in qualitative research

The suitability of the term 'reliability' for qualitative research is contested (e.g. Winter, 2000; Stenbacka, 2001; Golafshani, 2003). Lincoln and Guba (1985) prefer to replace 'reliability' with terms such as 'credibility', 'neutrality', 'confirmability', 'dependability', 'consistency', 'applicability', 'trustworthiness' and 'transferability', in particular the notion of 'dependability'.

LeCompte and Preissle (1993, p. 332) suggest that the canons of reliability for quantitative research may be unworkable for qualitative research. Quantitative research may strive for replication: if the same methods are used with the same sample then the results should be the same. Further, some quantitative methods require a degree of control and manipulation of phenomena. This distorts the natural occurrence of phenomena (see section above on 'Ecological validity'). Indeed the premises of naturalistic studies include the uniqueness and idiosyncrasy of situations, such that the study cannot be replicated; that is their strength rather than their weakness.

On the other hand, this is not to say that qualitative research need not strive for replication in generating, refining, comparing and validating constructs. Indeed LeCompte and Preissle (1993, p. 334) argue that such replication might include repeating:

- the status position of the researcher;
- the choice of informant/respondents;
- the social situations and conditions;
- the analytic constructs and premises that are used;
- the methods of data collection and analysis.

Further, Denzin and Lincoln (1994) suggest that reliability as replicability in qualitative research can be addressed in several ways:

- stability of observations (whether the researcher would have made the same observations and interpretation of these if they had been observed at a different time or in a different place);
- parallel forms (whether the researcher would have made the same observations and interpretations of what had been seen if she had paid attention to other phenomena during the observation);
- inter-rater reliability (whether another observer with the same theoretical framework and observing the same phenomena would have interpreted them in the same way).

This is a contentious issue, for it is seeking to apply to qualitative research the canons of reliability of quantitative research. Purists might argue against the legitimacy, relevance or need for this in qualitative studies.

In qualitative research, reliability can be regarded as a fit between what researchers record as data and what actually occurs in the natural setting that is being researched, i.e. a degree of accuracy and comprehensiveness of coverage (Bogdan and Biklen, 1992, p. 48). This is not to strive for uniformity: two researchers who are studying a single setting may come up with very different findings, but both sets of findings might be reliable. Indeed Kvale (1996, p. 181) suggests that there might be as many different interpretations of

qualitative data as there are researchers. An example of this is the study of the Nissan automobile factory in the UK, where Wickens (1987) found a 'virtuous circle' of work organization practices that demonstrated flexibility, teamwork and quality consciousness, whereas the same practices were reported by Garrahan and Stewart (1992) to be a 'vicious circle' of exploitation, surveillance and control respectively. Both versions of the same reality coexist because reality is not unitary. This argues for reliability to adopt an eclectic use of instruments, researchers, perspectives and interpretations (echoing the comments earlier about triangulation).

Brock-Utne (1996) argues that qualitative research, being holistic, strives to record the multiple interpretations of, intentions in and meanings given to situations and events. Here reliability is construed as *dependability* (Lincoln and Guba, 1985, pp. 108–9; Anfara *et al*., 2002), recalling the earlier discussion on internal validity. Dependability involves member checks (respondent validation), debriefing by peers, triangulation, prolonged engagement in the field, persistent observations in the field, reflexive journals, negative case analysis and independent audits (identifying acceptable processes of conducting the inquiry so that the results are consistent with the data). Audit trails enable the research to address the issue of confirmability of results, in terms of process and product (Golafshani, 2003, p. 601).

Dependability raises the important issue of *respondent validation* (researchers take back their research report to the respondents and record their reactions to that report). Whilst dependability might suggest that researchers should go back to respondents to check that their findings are dependable, researchers also need to be cautious in placing exclusive store on respondents, for, as Hammersley and Atkinson (1983) suggest, they are not in a privileged position to be sole commentators on their actions.

Kleven (1995) suggests that qualitative research can address reliability in part by asking three questions, particularly in observational research:

1. Would the same observations and interpretations have been made if observations had been conducted at different times? (The 'stability' version of reliability.)
2. Would the same observations and interpretations have been made if other observations had been conducted at the time? (The 'parallel forms' version of reliability.)
3. Would another observer, working in the same theoretical framework, have made the same observations and interpretations? (The 'inter-rater' version of reliability.)

The debate on reliability in quantitative and qualitative research rehearses the discussion of paradigms in the opening chapters: quantitative measures are criticized for combining sophistication and refinement of process with crudity of concept (Ruddock, 1981) and for failing to distinguish between educational and statistical significance (Eisner, 1985); qualitative methodologies, whilst possessing immediacy, flexibility, authenticity, richness and candour, are criticized for being impressionistic, biased, commonplace, insignificant, ungeneralizable, idiosyncratic, subjective and short-sighted (Ruddock, 1981). This is an arid debate; rather the issue is one of fitness for purpose. For our purposes here, we need to note that criteria of reliability in quantitative methodologies may differ from those in qualitative methodologies. In qualitative methodologies, reliability includes fidelity to real life, context- and situation-specificity, authenticity, comprehensiveness, detail, honesty, depth of response and meaningfulness to the respondents.

We summarize some similarities and differences between reliability in quantitative and qualitative research in Table 14.2.

Table 14.2 shows that, whilst there are some areas of reliability which are exclusive to quantitative research (split-half testing, equivalent forms and Cronbach alphas), many features of reliability apply, *mutatis mutandis*, to both quantitative and qualitative research. Further, Table 14.2 also shows that some features of validity (Table 14.1) also appear in reliability (e.g. content validity appears as coverage of domain and comprehensiveness, and concurrent validity appears as triangulation). This suggests some blurring of the edges between validity and reliability in the literature.

14.11 Validity and reliability in interviews

In interviews, inferences about validity are made too often on the basis of face validity (Cannell and Kahn, 1968), that is, whether the questions asked look as if they are measuring what they claim to measure. One cause of invalidity is bias, defined as 'a systematic or persistent tendency to make errors in the same direction, that is, to overstate or understate the "true value" of an attribute' (Lansing *et al*., 1961, pp. 120–1). One way of validating interview measures is to compare the interview measure with another measure that has already been shown to be valid, i.e. 'convergent validity', discussed earlier. If the two measures agree, it can be assumed that the validity of the interview is comparable with the proven validity of the other measure.

A practical way of achieving greater validity in interviews is to minimize bias as much as possible. Sources

TABLE 14.2 COMPARING RELIABILITY IN QUANTITATIVE AND QUALITATIVE RESEARCH

Bases of reliability in quantitative research		*Bases of reliability in qualitative research*
Reliability	⟷	Dependability
Demonstrability	⟷	Trustworthiness
Stability	⟷	Stability
Isolation, control and manipulation of required variables	⟷	Fidelity to the natural situation and real life
Identification, control and manipulation of key variables	⟷	Thick description and high detail on required or important aspects
Singular, objective truths	⟷	Multiple interpretations/perceptions
Replicability	⟷	Replicability
Parallel forms	⟷	Parallel forms
Generalizability	⟷	Generalizability
Context-freedom	⟷	Context-specificity
Objectivity	⟷	Authenticity
Coverage of domain	⟷	Comprehensiveness of situation
Verification of data and analysis	⟷	Honesty and candour
Answering research questions	⟷	Depth of response
Meaningfulness to the research	⟷	Meaningfulness to respondents
Parsimony	⟷	Richness
Objectivity	⟷	Confirmability
Fidelity to 'etic' research	⟷	Fidelity to 'emic' research
Internal consistency	⟷	Credibility
Generalizability	⟷	Transferability
Parallel forms	⟷	Parallel forms
Inter-rater reliability	⟷	Inter-rater reliability
Accuracy	⟷	Accuracy
Precision	⟷	Accuracy
Replication	⟷	Replication
Neutrality	⟷	Multiple interests represented
Consistency	⟷	Consistency
Theoretical relevance	⟷	Applicability
Triangulation	⟷	Triangulation
Alternative forms (equivalence)		
Split-half		
Inter-item correlations (alphas)		

of bias are: the characteristics of the interviewer and the respondent; and the substantive content of the questions. Researcher bias (Maxwell, 2005, p. 108), which has an effect on the interview (e.g. reactivity), can include, for example:

- the attitudes, opinions and expectations of the interviewer;
- a tendency for the interviewer to see the respondent in her own image;
- a tendency for the interviewer to seek answers that support her preconceived notions or theory;
- misperceptions on the part of the interviewer of what the respondent is saying;
- misunderstandings on the part of the respondent of what is being asked.

Studies have also shown that ethnicity, religion, gender, sexual orientation, status, social class and age in certain contexts can be potent sources of bias, i.e. interviewer effects (Lee, 1993; Scheurich, 1995). Interviewers and interviewees alike bring their own, often unconscious experiential and biographical baggage with them into the interview situation. Indeed Hitchcock and Hughes (1989) argue that because interviews are interpersonal, humans interacting with humans, it is inevitable that the researcher will have some influence on the interviewee and, thereby, on the data. Interviewer neutrality is a chimera (Denscombe, 1995, 2014).

Lee (1993) indicates problems of reliability in conducting interviews very sharply, where the researcher is researching sensitive subjects, i.e. research that might

pose a significant threat to interviewers and interviewees. Here the interview might be seen as an intrusion into private worlds, or the interviewer might be regarded as someone who can impose sanctions on the interviewee, or as someone who can exploit the powerless; the interviewee is in the interviewer's searchlight (see also Scheurich, 1995), so may be cautious in what is revealed. Indeed Gadd (2004) reports that an interviewee may reduce his/her willingness to 'open up' to an interviewer if the dynamics of the interview situation are too threatening, taking the role of the 'defended subject'. This raises issues of *transference* and *counter-transference*, which have their basis in psychoanalysis. In transference, interviewees project onto the interviewer their feelings, fears, desires, needs and attitudes that derive from their own experiences (Scheurich, 1995). In counter-transference the process is reversed. Both affect reliability.

One way of addressing reliability is to have a highly structured interview, with the same format and sequence of words and questions for each respondent (Silverman, 1993), though Scheurich (1995, pp. 241–9) suggests that this is to misread the infinite complexity and open-endedness of social interaction. Controlling the wording is no guarantee of controlling the interview. Oppenheim (1992, p. 147) argues that wording is a particularly important factor in attitudinal rather than factual questions. He suggests that changes in wording, context and emphasis undermine reliability, because it ceases to be the same question for each respondent. Indeed he argues that error and bias can stem from alterations to wording, procedure, sequence, recording and rapport, and that training for interviewers is essential to minimize this. Silverman (1993) suggests that it is important for each interviewee to understand the question in the same way. He suggests that the reliability of interviews can be enhanced by: careful piloting of interview schedules; training of interviewers; inter-rater reliability in the coding of responses; and the extended use of closed questions.

On the other hand, Silverman (1993) argues for the importance of open-ended interviews, as this enables respondents to demonstrate their unique way of looking at the world – their definition of the situation. It recognizes that what is a suitable sequence of questions for one respondent might be less suitable for another, and open-ended questions enable important but unanticipated issues to be raised.

Oppenheim (1992, pp. 96–7) suggests several causes of bias in interviewing:

- biased sampling (sometimes created by the researcher not adhering to sampling instructions);
- poor rapport between interviewer and interviewee;
- changes to question wording (e.g. in attitudinal and factual questions);
- poor prompting and biased probing;
- poor use and management of support materials (e.g. show cards);
- alterations to the sequence of questions;
- inconsistent coding of responses;
- selective or interpreted recording of data/transcripts;
- poor handling of difficult interviews.

One can add to this the issue of 'acquiescence' (Breakwell, 2000, p. 254), the tendency of respondents to say 'yes', regardless of the question or, indeed, regardless of what they really feel or think.

There is also the issue of *leading questions*. A leading question is one which makes assumptions about interviewees or 'puts words into their mouths', i.e. where the question influences the answer perhaps illegitimately. For example (Morrison, 1993, pp. 66–7), the question 'when did you stop complaining to the headteacher?' assumes that the interviewee had been a frequent complainer, and the question 'how satisfied are you with the new Mathematics scheme?' assumes a degree of satisfaction with the scheme. The leading questions here might be rendered less leading by rephrasing, for example: 'how frequently do you have conversations with the headteacher?' and 'what is your opinion of the new Mathematics scheme?' respectively.

In discussing the issue of leading questions we are not necessarily suggesting that there is not a place for them. Indeed Kvale (1996, p. 158) makes a powerful case *for* leading questions, arguing that they may be necessary in order to obtain information that the interviewer suspects the interviewee might be withholding. Here it might be important to put the 'burden of denial' onto the interviewee (e.g. 'when did you stop cheating in examinations?'). Leading questions, frequently used in police interviews, may be used for reliability checks with what the interviewee has already said, or may be deliberately used to elicit particular non-verbal behaviours that provide an indication of the sensitivity of the interviewee's remarks.

The researcher must also be aware of possible bias in interviewees giving what they think are socially desirable answers to questions (Fowler, 2009), i.e. answers to please the interviewer or not to appear different from what is socially acceptable or desirable.

Reducing bias in interviews requires: (a) careful formulation of questions so that the meaning is crystal clear; (b) thorough training procedures so that an interviewer is more aware of the possible problems; (c) probability sampling of respondents; and (d) matching

interviewer characteristics with those of the sample being interviewed (where appropriate). Oppenheim (1992, p. 148) argues, for example, that interviewers seeking attitudinal responses have to ensure that people with known characteristics are included in the sample – the criterion group. Researchers must recognize that the interview is a shared, negotiated and dynamic social moment.

Power is significant in the interview situation, for the interview is not simply a data-collection situation but a social and frequently a political situation. Literally the word 'inter-view' is a view *between* people, mutually, not the interviewer alone, extracting data one way from the interviewee. Power resides with interviewer and interviewee alike (Thapar-Björkert and Henry, 2004), though Scheurich (1995, p. 246) argues that, typically, more power resides with the interviewer (see also Lee, 1993; Morrison, 2013a): the interviewer generates the questions and the interviewee answers them; the interviewee is under scrutiny whilst the interviewer is not. Kvale (1996, p. 126), too, suggests that there are definite asymmetries of power as the interviewer tends to define the situation, the topics and the course of the interview. Of course, the interviewee is powerful as he/she has data that the interviewer wants, and has power to withhold such data (discussed below).

Cassell (cited in Lee, 1993) and Walford (2012) suggest that elites and powerful people might feel demeaned or insulted when being interviewed by those with a lower status or less power. Further, those with power, resources and expertise might be anxious to maintain their reputation, and so will be more guarded in what they say, wrapping this up in well-chosen, articulate phrases (Walford, 2012). Interviewers need to be aware of the potentially distorting effects of power, a significant feature of critical theory (see Chapter 3).

Neal (1995) draws attention to the feelings of powerlessness and anxieties about the physical presentation and status of interviewers when interviewing powerful people. This is particularly so for frequently lone, low-status research students interviewing powerful people; a low-status female research student might find that an interview with a male in a position of power (e.g. a university vice-chancellor, a senior politician or a senior manager) might turn out to be very different from an interview with the same person if conducted by a male university professor, which is perceived by the interviewee to be more of a dialogue between equals (see also Connell *et al*., 1996; Gewirtz and Ozga, 1993, 1994). Ball (1994b) comments that, when powerful people are being interviewed, interviews must be seen as an extension of the 'play of power' – with its game-like connotations. He suggests that powerful people control the agenda and course of the interview, and are usually very adept at this because they have both a personal and professional investment in being interviewed (see also Batteson and Ball, 1995; Phillips, 1998; Walford, 2012).

The effect of power can be felt even before the interview commences. Neal (1995) instances being kept waiting, and subsequently being interrupted, patronized and interviewed by the interviewee (see also Walford, 1994b, 2012). Scheurich (1995) suggests that many powerful interviewees will rephrase or not answer the question. Limerick *et al*. (1996) report interviews in which interviewers have felt themselves to be passive, vulnerable, helpless and indeed manipulated. One way of overcoming this is to have two interviewers conducting each interview (Walford, 1994c, p. 227). On the other hand, Hitchcock and Hughes (1989) observe that if the researchers are known to the interviewees and they are peers, however powerful, then a degree of reciprocity might be taking place, with interviewees giving answers that they think the researchers might want to hear.

The issue of power features in feminist research (e.g. Thapar-Björkert and Henry, 2004), i.e. research which emphasizes subjectivity, equality, reciprocity, collaboration, non-hierarchical relations and emancipatory potential (catalytic and consequential validity) (Neal, 1995), echoing the comments on research that is influenced by critical theory (Chapter 3). Here feminist research addresses a dilemma of interviews which are constructed in the dominant, male paradigm of pitching questions that demand answers from a passive respondent.

Limerick *et al*. (1996) suggest that, in fact, it is wiser to regard the interview as a gift (and 'data' means 'things that are given'), as interviewees have the power to withhold information, to choose the location of the interview, to choose how seriously to attend to the interview, how long it will last, when it will take place, what will be discussed – and in what and whose terms – what knowledge is important, even how the data will be analysed and used (see also Thapar-Björkert and Henry, 2004). Echoing Foucault, they argue that power is fluid and is discursively constructed through the interview rather than being the province of either party.

Miller and Cannell (1997) identify particular problems in conducting telephone interviews, where reducing the interview situation to just auditory sensory cues can be challenging (see Chapter 25). For example, the interviewee can only retain a certain amount of information in her/his short-term memory, so bombarding the telephone interviewee with too many choices (the

non-written form of 'show cards' of possible responses) becomes unworkable. Here the reliability of responses is subject to the memory capabilities of the interviewee – how many scale points and descriptors, for example, can an interviewee retain in her head about a single item? Further, the absence of non-verbal cues, for example, facial expression, gestures, posture, the significance of silences and pauses (Robinson, 1982), is important, as interviewees may be unclear about the meaning behind words and statements. This problem is compounded if the interviewer is unknown to the interviewee.

Miller and Cannell (1997) report important research evidence to support the significance of the non-verbal mediation of verbal dialogue. As discussed earlier, the interview is a social situation; in telephone interviews the absence of essential social elements could undermine the salient conduct of the interview, and hence its reliability and validity. Non-verbal paralinguistic cues affect the conduct, pacing and relationships in the interview and the support, threat and confidence felt by the interviewees. Telephone interviews can easily slide into becoming mechanical and cold. Further, the problem of loss of non-verbal cues is compounded by the asymmetries of power that often exist between interviewer and interviewee; the interviewer will need to take immediate steps to address these issues (e.g. by putting interviewees at their ease, as the interviewee might simply put down the telephone).

On the other hand, Nias (1991), Miller and Cannell (1997) and Ary *et al*. (2002) suggest that the very fact that interviews are not face-to-face may strengthen their reliability, as the interviewee might disclose information that may not be so readily forthcoming in a face-to-face, more intimate situation. Hence, telephone interviews have their strengths and weaknesses; their use should be governed by the criterion of fitness-for-purpose. They tend to be shorter, more focused and useful for contacting busy people (Harvey, 1988; Miller, 1995).

A cluster of problems surround the person being interviewed. Tuckman (1972), for example, observed that, when formulating questions, an interviewer has to consider the extent to which a question might influence the respondent to show herself in a good light; or the extent to which a question might influence the respondent to be unduly helpful by attempting to anticipate what the interviewer wants to hear; or the extent to which a question might be asking for information about a respondent that she is not certain or likely to know herself. Interviews may be based on the assumption that the person interviewed has insight into the cause of her behaviour, and this may not be possible.

In educational circles interviewing might be a particular problem in working with children (Morrison, 2013a) (see also Chapters 13 and 25). Simons (1982), McCormick and James (1988) and Greig and Taylor (1999) comment on particular problems involved in interviewing children which affect reliability, for example:

- establishing trust;
- overcoming shyness and reticence;
- maintaining informality;
- avoiding assuming that children 'know the answers';
- overcoming the problems of inarticulate children;
- pitching the question at the right level;
- choice of vocabulary;
- non-verbal cues;
- moving beyond the institutional response or receiving what children think the interviewer wants to hear;
- avoiding the interviewer being seen an authority spy or plant;
- keeping to the point;
- breaking silences on taboo areas and those which are reinforced by peer-group pressure;
- children being seen as of lesser importance than adults (maybe in the sequence in which interviews are conducted, e.g. the headteacher, then the teaching staff, then the children).

These are not new matters. Studies by Labov in the 1960s and 1970s showed how students reacted very strongly to contextual matters in an interview situation (Labov, 1969). The language of children varied according to the ethnicity of the interviewee, the friendliness of the surroundings, the opportunity for the children to be interviewed with friends, the ease with which the scene was set for the interview, the demeanour of the adult (e.g. whether the adults was standing or sitting), the nature of the topics covered. The interview is a social encounter, and children may be very sensitive to the social context of the interview (Morison *et al*., 2000; Morrison, 2013a); Maguire (2005, p. 4) suggests that 'children have good social radar'. The differences can be significant, varying from monosyllabic responses by children in unfamiliar and uncongenial surroundings to extended responses in the more congenial and less threatening surroundings – more sympathetic to the children's everyday world. The language, argot and jargon, social and cultural factors of the interviewer and interviewee all exert a powerful influence on the interview situation.

Lee (1993) raises the further issue of whether there should be a single interview that maintains the detachment of the researcher (perhaps particularly useful in

addressing sensitive topics), or whether there should be repeated interviews to gain depth and to show fidelity to the collaborative nature of research (a feature, as noted above, which is significant for feminist research (Oakley, 1981)).

Kvale (1996, pp. 148–9) suggests that, in order to obtain reliable and valid data, a skilled interviewer should:

- know his/her subject matter in order to conduct an informed conversation;
- structure the interview well, so that each stage of the interview is clear to the participant;
- be clear in the terminology and coverage of the material;
- allow participants to take their time and answer in their own way;
- be sensitive and empathic, using active listening and being sensitive to how something is said and the non-verbal communication involved;
- be alert to those aspects of the interview which may hold significance for the participant;
- keep to the point and the matter in hand, steering the interview where necessary in order to address this;
- check the reliability, validity and consistency of responses by well-placed questioning;
- be able to recall and refer to earlier statements made by the participant;
- be ready to clarify, confirm and modify the participant's comments with the participant.

Walford (1994c, 2012) adds to this the need for the interviewer to have done her homework when interviewing powerful people, as such people could well interrogate the interviewer – they will assume up-to-dateness, competence and knowledge in the interviewer. Powerful interviewees are usually busy people and will expect the interviewer to have read relevant material in the public domain.

The issues of reliability do not reside solely in the preparations for and conduct of the interview; they extend to the ways in which interview data are analysed. For example, Lee (1993) and Kvale (1996, p. 163) comment on the issue of 'transcriber selectivity'. Here transcripts of interviews, however detailed and full they might be, remain selective, since they are interpretations of social situations. They become decontextualized, abstracted, even if they record silences, intonation, non-verbal behaviour etc. The issue, then, is how useful they are to researchers overall rather than whether they are completely reliable.

One problem in using open-ended questions in interviews is that of developing a satisfactory, reliable method of recording replies. One way is to summarize responses in the course of the interview. This has the disadvantage of breaking the continuity of the interview and may result in bias because the interviewer may unconsciously emphasize responses that agree with her expectations and fail to note those that do not. It is sometimes possible to summarize an individual's responses at the end of the interview. Although this preserves the continuity of the interview, it is likely to induce greater bias because the delay may lead to the interviewer forgetting some of the details. It is these forgotten details that are most likely to be those which disagree with his or her own expectations. We advise the reader also to review Chapter 25 of the present volume.

14.12 Validity and reliability in experiments

One fundamental purpose of experimental design is to impose control over conditions that would otherwise cloud the true effects of the independent variables on the dependent variables, so that causality can be attributed to the intervention in question. Clouding conditions that threaten to jeopardize the validity of experiments have been identified by Campbell and Stanley (1963), Bracht and Glass (1968), Lewis-Beck (1993), Shadish *et al*. (2002) and Torgerson and Torgerson (2008), conditions that are of greater consequence to the validity of quasi-experiments (more typical in educational research) than to true experiments in which random sampling and assignment to treatments occurs and where both treatment and measurement can be more adequately controlled by the researcher.

The following summaries distinguish between 'internal validity' and 'external validity'. Internal validity is concerned with the question, do the experimental treatments, in fact, make a difference in the specific experiments under scrutiny? External validity, on the other hand, asks the question, given these demonstrable effects, to what populations or settings can they be generalized?

Threats to internal validity were introduced earlier in this chapter, and comprise:

- history
- maturation
- statistical regression
- testing
- instrumentation
- selection
- experimental mortality
- instrument reactivity
- selection–maturation interaction.

Shadish *et al*. (2002), Torgerson and Torgerson (2008) and Creswell (2012) add to these: (a) contamination, in that control and experimental groups may communicate with each other, affecting what happens with each group; and (b) compensatory rivalry: a control group may feel resentful about being deprived of the intervention (if they are told what the intervention comprises) and this may affect their behaviour.

Several threats to external validity were discussed earlier in this chapter (Section 14.5) and the reader is advised to review these. As in clinical trials, educational experiments require attention to the educational equivalents of: effects on different sub-groups of a sample; side effects; contra-indications; effects of personal characteristics of participants; dose-response (e.g. attention to amount, quality, strength, frequency, intensity, duration of the intervention); recognition that a person is a complex system which combines and connects very many elements whose interactions and outcomes change over time (with commensurate changes to interventions over time); patients who forget to, or refuse to, take medicine or take it irregularly (i.e. changes to the intervention affect reliability); comorbidity (other problems may exist at the time of the intervention); patients taking multiple medications (several other events may be happening at the same time as the intervention); and doctor–patient (teacher–student) relations.

In clinical trials, some treatments may require initial intensive medication, tailing off to a maintenance level (e.g. many people start antibiotics with a double dose and then maintain a single dose for several days); some medication may require a gentle start, with an increasing, cumulative dosage as the body responds to treatment. The equivalents in educational research also apply. Further, some drugs may require ongoing, regular, very close monitoring, whilst others require less frequent monitoring. The point here is that medication is not a single event but, as the career of a disease and a patient changes over time, so does the treatment. Translating these practices from clinical research to educational research suggests that educational experiments need greater sensitivity to many confounding factors and to the need to address many kinds of threats to validity and reliability.

Though the analogy between clinical trials and educational experiments may not hold too strongly, for example, the former operating on a one-to-one patient–doctor relation and the latter typically operating on a one-to-many basis, and the former adopting a pathological model and the latter adopting a non-pathological model, nevertheless in educational experiments, this argues against a too-simplistic operation of randomized controlled trials, with its focus on average effects and its risk of overlooking the importance of outliers and sub-groups. This suggests that, to ensure validity and reliability in educational experiments, attention must be paid to: the whole person; contexts and other interventions and practices that are taking place in the educational experiences of students at the time of the experiment; duration, intensity and differential inputs into, operations of and responses to the intervention; and to the often non-linear and dynamical systems nature of the 'careers' of participants.

An experiment can be said to be internally valid to the extent that, within its own confines, its results are credible; but for those results to be useful, they must be generalizable beyond the confines of the particular experiment; in a word, they must be externally valid also (for a critique of randomized controlled experiments and the problems of generalizability, see Morrison, 2001; Cartwright and Hardie, 2012). Without internal validity an experiment cannot be externally valid, but the converse does not necessarily follow; an internally valid experiment may or may not have external validity.

It follows, then, that the way to reliable and valid educational experimentation lies in maximizing both internal validity and, where relevant, external validity.

14.13 Validity and reliability in questionnaires

Questionnaires feature in many types of research, from surveys to action research. Validity of questionnaires (however administered, e.g. face-to-face, postal, telephone, Internet) can be seen from two viewpoints (Belson, 1986). First, whether respondents who complete questionnaires do so accurately, honestly and correctly; and second, whether those who fail to return their questionnaires would have given the same distribution of answers as did the returnees.

As more and more questionnaires are conducted online, this brings with it the problem of honesty: are respondents really telling the truth about themselves and about the matters to which they have been asked to respond? Fowler (2009) gives the example of people under-reporting how many cigarettes they smoke each day or how much alcohol they drink (p. 16). Fowler also reports that respondents may not understand or may misunderstand a question, or they may not know the answer, or they may not be able to recall accurately, or they may not want to disclose information, or they may give what they believe to be the socially desirable answer rather than a 'true' answer, and all of these affect reliability (p. 105). Here the piloting of the

questionnaire and the guarantees of anonymity and non-traceability might attenuate such problems.

The question of accuracy can be checked by the intensive interview method, a technique consisting of tactics that include familiarization, temporal reconstruction, probing and challenging (cf. Belson, 1986, pp. 35–8).

The problem of non-response is addressed in Chapters 17 and 40. Hudson and Miller (1997) suggest several strategies for maximizing the response rate to questionnaires (and, thereby, to increase reliability), which involve:

- including stamped addressed envelopes (for postal questionnaires);
- multiple rounds of follow-up to request returns (maybe up to three follow-ups);
- stressing the importance and benefits of the questionnaire;
- stressing the importance of, and benefits to, the client group being targeted (particularly if it is a minority group that is struggling to have a voice);
- providing interim data from returns to non-returners to involve and engage them in the research;
- checking addresses and changing them if necessary;
- following up questionnaires with a personal telephone call;
- tailoring follow-up requests to individuals (with indications to them that they are personally known and/or important to the research – including providing respondents with clues by giving some personal information to show that they are known) rather than blanket generalized letters;
- features of the questionnaire itself (ease of completion, time to be spent, sensitivity of the questions asked, length of the questionnaire);
- invitations to a follow-up interview (face-to-face or by telephone);
- encouragement to participate by a friendly third party;
- understanding the nature of the sample population in depth, so that effective targeting strategies can be used.

The advantages of the questionnaire over interviews, for instance, are: it tends to be more reliable; because it is anonymous, it may encourage greater honesty (though, of course, dishonesty and falsification might not be able to be discovered in a questionnaire, and this is a particular issue in Internet questionnaires, where even the factual details about a respondent may be false). For example, an online respondent can create a different persona and make up responses (Shulman *et al*., 2011; Ramírez and Palu-ay, 2015). Further, a questionnaire is often more economical than the interview in terms of time and money, for example, it can be mailed or conducted online.

Its disadvantages, on the other hand, are: there is often too low a percentage of returns; the interviewer is unable to answer questions concerning both the purpose of the interview and any misunderstandings experienced by the interviewee, for it sometimes happens in the case of the latter that the same questions have different meanings for different people; if only closed items are used, the questionnaire may lack coverage or authenticity; if only open items are used, respondents may be unwilling to write their answers for one reason or another; questionnaires present problems to people of limited literacy; and an interview can be conducted at an appropriate speed whereas questionnaires are often filled in hurriedly. There is a need, therefore, to pilot questionnaires and refine their contents, wording, length, etc. as appropriate for the sample being targeted.

One central issue in considering the reliability and validity of questionnaire surveys is that of sampling. An unrepresentative, skewed sample, one that is too small or too large, can easily distort the data, and indeed, in the case of very small samples, prohibit statistical analysis. We address sampling in Chapter 12.

14.14 Validity and reliability in observations

There are questions about two types of validity in observation-based research. Comments about the subjective and idiosyncratic nature of the participant observation study are about its external validity. How do we know that the results of this one piece of research are applicable to other situations? Fears that observers' judgements will be affected by their close involvement in the group relate to the internal validity of the method. How do we know that the results of this one piece of research represent the real thing? In Chapter 12 on sampling, we refer to a number of techniques (quota sampling, snowball sampling, purposive sampling) that researchers employ as a way of checking on the representativeness of the events that they observe and of cross-checking their interpretations of the meanings of those events.

In addition to external validity, participant observation should have rigorous internal validity checks. There are several threats to validity and reliability here, for example:

- the researcher, in exploring the present, may be unaware of important antecedent events;

- informants may be unrepresentative of the sample in the study;
- the presence of the observer might bring about different behaviours (reactivity);
- the researcher might 'go native', becoming too attached to the group to see it sufficiently dispassionately.

To address this, Denzin (1989) suggests triangulation of data sources and methodologies. Chapter 26 discusses the principal ways of overcoming problems of reliability and validity in observational research in naturalistic inquiry. In essence it is suggested that 'trustworthiness' (Lincoln and Guba, 1985) replaces more conventional views of reliability and validity, and that this is devolved on issues of *credibility*, *confirmability*, *transferability* and *dependability*.

If observational research is more structured in its nature, yielding quantitative data, then the conventions of intra- and inter-rater reliability apply. Here steps are taken to ensure that observers enter data into the appropriate categories consistently (i.e. intra- and inter-rater reliability) and accurately. Further, to ensure validity, a pilot should have been conducted to ensure that the observational categories themselves are appropriate, exhaustive, discrete, unambiguous and effectively operationalize the purposes of the research.

14.15 Validity and reliability in tests

The researcher will have to judge the place and significance of test data, not forgetting the problem of the Hawthorne effect operating negatively or positively on students who undertake the tests. There is a range of issues which might affect the reliability of the test – for example, the time of day, the time of the school year, the temperature in the test room, the perceived importance of the test, the degree of formality of the test situation, 'examination nerves', the amount of guessing of answers by the students, how the test is administered and marked, the degree of closure or openness of test items. Hence the researcher who is considering using testing for acquiring research data must ensure that it is appropriate, valid and reliable (Linn, 1993; Borsboom *et al*., 2004).

Wolf (1994) suggests four main factors that might affect reliability: the range of the group that is being tested; the group's level of proficiency; the length of the measure (the longer the test, the greater the chance of errors); and the way in which reliability is calculated. (Fitz-Gibbon (1997, p. 36) argues that, *ceteris paribus*, longer tests are more reliable than shorter tests.) Additionally, there are several ways in which reliability might be compromised in tests. Feldt and Brennan (1993) suggest four types of threat to reliability:

- individuals (e.g. their motivation, concentration, forgetfulness, health, carelessness, guessing, their related skills, e.g. reading ability, their usedness to solving the type of problem set, the effects of practice);
- situational factors (e.g. the psychological and physical conditions for the test – the context);
- test marker factors (e.g. idiosyncrasy and subjectivity);
- instrument variables (e.g. poor domain sampling, errors in sampling tasks, the realism of the tasks and relatedness to the experience of the testees, poor question items, the assumption or extent of unidimensionality in item response theory, length of the test, mechanical errors, scoring errors, computer errors).

Sources of unreliability

There are several threats to reliability in tests and examinations, particularly tests of performance and achievement, for example (Cunningham, 1998; Airasian, 2001; Cohen *et al*., 2010; Creswell, 2012).

For example, with respect to *examiners* and *markers* these concern:

- errors in marking (e.g. attributing, adding and transfer of marks);
- inter-rater reliability (different markers giving different marks for the same or similar pieces of work);
- inconsistency in the marker (e.g. being harsh in the early stages of the marking and lenient in the later stages of the marking of many scripts);
- variations in the award of grades for work that is close to grade boundaries (some markers placing the score in a higher or lower category than other markers);
- the Halo effect, wherein a student who is judged to do well or badly in one assessment is given undeserved favourable or unfavourable assessment respectively in other areas.

With reference to the *students* and *teachers* themselves, there are several sources of unreliability:

- Motivation and interest in the task has a considerable effect on performance. Clearly, students need to be motivated if they are going to make a serious attempt at any test that they are required to

undertake, where motivation is intrinsic (doing something for its own sake) or extrinsic (doing something for an external reason, e.g. obtaining a certificate or employment or entry into higher education). The results of a test completed in a desultory fashion by resentful pupils are hardly likely to supply the researcher with reliable information about the students' capabilities (Wiggins, 1998). Motivation to participate in test-taking sessions is strongest when students have been helped to see its purpose, and where the examiner maintains a warm, purposeful attitude towards them during the testing session (Airasian, 2001).

- The relationship (positive to negative) between the assessor and the testee exerts an influence on the assessment. This takes on increasing significance in teacher assessment, where the students know the teachers personally and professionally – and vice versa – and where the assessment situation involves face-to-face contact between the teacher and the student. Both test-takers and test-givers mutually influence one another during examinations, oral assessments and the like (Harlen, 1994). During a face-to-face test, students respond to such characteristics of the assessor as the person's sex, age and personality.
- The conditions – physical, emotional, social – exert an influence on the assessment, particularly if they are unfamiliar. Wherever possible, students should take tests in familiar settings, preferably in their own classrooms under normal school conditions. Distractions in the form of extraneous noise, walking about the room by the examiner, and intrusions into the room, all have significant impact upon the scores of the test-takers, particularly when they are younger pupils (Gipps, 1994). An important factor in reducing students' anxiety and tension during an examination is the extent to which they are quite clear about what exactly they are required to do. Simple instructions, clearly and calmly given by the examiner, can significantly lower the general level of tension in the test-room. Teachers who intend to conduct testing sessions may find it beneficial in this respect to rehearse the instructions they wish to give to pupils before the actual testing session. Ideally, test instructions should be simple, direct and as brief as possible.
- The Hawthorne effect, wherein, in this context, simply informing a student that this is an assessment situation will be enough to disturb her performance – for better or worse (either case not being a fair reflection of her usual abilities).
- Distractions (including superfluous information).
- Students respond to the tester in terms of their perceptions of what he/she expects of them (Haladyna, 1997; Tombari and Borich, 1999; Stiggins, 2001).
- The time of the day, week or month will exert an influence on performance. Some students are fresher in the morning and more capable of concentration (Stiggins, 2001).
- Students are not always clear on what they think is being asked in the question; they may know the right answer but not infer that this is what is required in the question.
- The students may vary from one question to another – a student may have performed better with a different set of questions which tested the same matters. Black (1998) argues that two questions which, to the expert, may seem to be asking the same thing but in different ways, to the students might well be seen as completely different questions.
- Students (and teachers) practice test-like materials, which, even though scores are raised, might make them better at taking tests, but the results might not indicate increased performance.
- A student may be able to perform a specific skill in a test but not be able to select or perform it in the wider context of learning.
- Cultural, ethnic and gender background affect how meaningful an assessment task or activity is to students, and meaningfulness affects their performance.
- Students' personalities may make a difference to their test performance.
- Students' learning strategies and styles may make a difference to their test performance.
- Marking practices are not always reliable, markers maybe being too generous, marking by effort and ability rather than performance.
- The context in which the task is presented affects performance: some students can perform the task in everyday life but not under test conditions.

With regard to the *test items* themselves, there may be problems (e.g. test bias), for example:

- The task itself may be multi-dimensional, for example, testing 'reading' may require several components and constructs. Students can execute a mathematics operation in the mathematics class but they cannot perform the same operation in, for example, a physics class; students will disregard English grammar in a science class but observe it in an English class. This raises the issue of the number of contexts in which the behaviour must be

demonstrated before a criterion is deemed to have been achieved (Cohen *et al*., 2010). The question of transferability of knowledge and skills is also raised in this connection. The context of the task affects the student's performance.

- The validity of the items may be in question.
- Clarity of the test items (Creswell, 2012): the avoidance of ambiguity.
- The language of the assessment and the assessor exerts an influence on the testee, for example, if the assessment is carried out in the testee's second language or in a 'middle-class' code (Haladyna, 1997).
- The readability level of the task can exert an influence on the test, for example, a difficulty in reading might distract from the purpose of a test which is of the use of a mathematical algorithm.
- The size and complexity of numbers or operations in a test (e.g. of mathematics) might distract the testee who actually understands the operations and concepts.
- The number and type of operations and stages to a task – a student might know how to perform each element, but when they are presented in combination the size of the task can be overwhelming.
- The form and presentation of questions affect the results, giving variability in students' performances.
- A single error early on in a complex sequence may confound the later stages of the sequence (within a question or across a set of questions), even though the student might have been able to perform the later stages of the sequence, thereby preventing the student from gaining credit for all she or he can, in fact, do.
- Questions might favour boys more than girls or vice versa.
- Essay questions favour boys if they concern impersonal topics and girls if they concern personal and interpersonal topics (Haladyna, 1997; Wedeen *et al.*, 2002).
- Boys may perform better than girls on multiple-choice questions and girls perform better than boys on essay-type questions (perhaps because boys are more willing than girls to guess in multiple-choice items), and girls may perform better in written work than boys.
- Questions and assessment may be culture-bound: what is comprehensible in one culture may be incomprehensible in another.
- The test may be so long, in order to ensure coverage, that boredom and loss of concentration may impair reliability.

With regard to the *operational procedures* of the test, there may be variability in:

- the conditions operating at the time of the test (e.g. noise, distractions, ambient, temperature, time of day/week);
- the test administration, with unstandardized procedure in the timing, duration, teacher intervention, instructions given, distribution and collection of materials/test papers, monitoring etc.

Hence specific contextual factors can exert a significant influence on learning and this has to be recognized in conducting assessments, to render an assessment as unthreatening and natural as possible.

Harlen (1994, pp. 140–2) suggests that inconsistency and unreliability in teacher- and school-based assessment may derive from differences in: (a) interpreting the assessment purposes, tasks and contents, by teachers or assessors; (b) the actual task set, or the contexts and circumstances surrounding the tasks (e.g. time and place); (c) how much help is given to the test-takers during the test; (d) the degree of specificity in the marking criteria; (e) the application of the marking criteria and the grading or marking system that accompanies it; (f) how much additional information about the student or situation is being referred to in the assessment.

Harlen advocates the use of a range of moderation strategies, both before and after the tests, including:

- statistical reference/scaling tests;
- inspection of samples (by post or by visit);
- group moderation of grades;
- post hoc adjustment of marks;
- accreditation of institutions;
- visits of verifiers;
- agreement panels;
- defining marking criteria;
- exemplification;
- group moderation meetings.

Whilst moderation procedures are essentially post hoc adjustments to scores, agreement trials and practice marking can be undertaken *before* the administration of a test, which is particularly important if there are large numbers of scripts or several markers.

The issue here is that results as well as instruments should be reliable. Reliability is also addressed by:

- calculating coefficients of reliability, split-half techniques, the Kuder-Richardson formula, parallel/equivalent forms of a test, test/re-test methods, the alpha coefficient;
- calculating and controlling the standard error of measurement;

- increasing the sample size (to maximize the range and spread of scores in a norm-referenced test), though criterion-referenced tests recognize that scores may bunch around the high level (in mastery learning, for example), i.e. the range of scores might be limited, thereby lowering the correlation coefficients that can be calculated;
- increasing the number of observations made and items included in the test (in order to increase the range of scores);
- ensuring effective domain sampling of items in tests based on item response theory (a particular issue in Computer adaptive testing, introduced below (Thissen, 1990));
- ensuring effective levels of item discriminability and item difficulty.

Reliability not only has to be achieved but has to be seen to be achieved, particularly in 'high-stakes' testing (where a lot hangs on the results of the test, e.g. entrance to higher education or employment). Hence the procedures for ensuring reliability must be transparent. The difficulty here is that the more one moves towards reliability as defined above, the more the test will become objective, the more students will be measured as though they are standardized objects, and the more the test will become decontextualized.

An alternative form of reliability which is premised on a more constructivist psychology emphasizes the significance of context, the importance of subjectivity and the need to engage and involve the testee more fully than a simple test. This rehearses the tension between positivism, post-positivism and more interpretive approaches outlined in the first chapter of this book. Objective tests, as described in the present chapter, lean strongly towards the positivist/post-positivist paradigm, whilst more phenomenological and interpretive paradigms of social science research will emphasize the importance of settings, of individual perceptions, of attitudes, in short, of 'authentic' testing (e.g. by using non-contrived, non-artificial forms of test data, e.g. portfolios, documents, course work, tasks that are stronger in realism and more 'hands on'). Though this latter adopts a view which is closer to assessment rather than narrowly 'testing', nevertheless the two overlap, both can yield marks, grades and awards, both can be formative as well as summative, both can be criterion-referenced.

With regard to validity, it is important to note here that an effective test will ensure adequate:

- content validity (e.g. adequate and representative coverage of programme and test objectives in the test items, a key feature of domain sampling); content validity is achieved by ensuring that the content of the test fairly samples the class or fields of the situations or subject matter in question. Content validity is achieved by making professional judgements about the relevance and sampling of the contents of the test to a particular domain. It is concerned with coverage and representativeness rather than with patterns of response or scores. It is a matter of judgement rather than measurement (Kerlinger, 1986). Content validity will need to ensure several features of a test (Wolf, 1994): (a) test coverage (the extent to which the test covers the relevant field); (b) test relevance (the extent to which the test items are taught through, or are relevant to, a particular programme); (c) programme coverage (the extent to which the programme covers the overall field in question);
- criterion-related validity (where a high correlation coefficient exists between the scores on the test and the scores on other accepted tests of the same performance); criterion-related validity is achieved by comparing the scores on the test with one or more variables (criteria) from other measures or tests that are considered to measure the same factor. Wolf (1994) argues that a major problem facing test devisers addressing criterion-related validity is the selection of the suitable criterion measure. He cites the example of the difficulty of selecting a suitable criterion of academic achievement in a test of academic aptitude. The criterion must be: (a) relevant (and agreed to be relevant); (b) free from bias (i.e. where external factors that might contaminate the criterion are removed); (c) reliable – precise and accurate; (d) capable of being measured or achieved;
- construct validity (e.g. the clear relatedness of a test item to its proposed construct/unobservable quality or trait, demonstrated by both empirical data and logical analysis and debate, i.e. the extent to which particular constructs or concepts can account for performance on the test); construct validity is achieved by ensuring that performance on the test is fairly explained by particular appropriate constructs or concepts. As with content validity, it is not based on test scores, but is more a matter of whether the test items are indicators of the underlying, latent construct in question. In this respect construct validity also subsumes content and criterion-related validity. Construct validity is threatened by (a) under-representation of the construct, i.e. the test is too narrow and neglects significant facets of a construct, and (b) the inclusion of irrelevancies – excess variance;

- concurrent validity (where the results of the test concur with results on other tests or instruments that are testing/assessing the same construct/performance – similar to predictive validity but without the time dimension. Concurrent validity can occur simultaneously with another instrument rather than after some time has elapsed);
- face validity (that, superficially, the test appears – at face value – to test what it is designed to test);
- jury validity (an important element in construct validity, where it is important to agree on the conceptions and operationalization of an unobservable construct);
- predictive validity (where results on a test accurately predict subsequent performance – akin to criterion-related validity);
- consequential validity (where the inferences that can be made from a test are sound);
- systemic validity (Fredericksen and Collins, 1989) (where programme activities both enhance test performance and enhance performance of the construct that is being addressed in the objective). Cunningham (1998) gives an example of systemic validity where, if the test and the objective of vocabulary performance leads to testees increasing their vocabulary, then systemic validity has been addressed.

To ensure test validity, then, the test must demonstrate fitness for purpose as well as addressing the several types of validity outlined above. The most difficult for researchers to address, perhaps, is construct validity, as it argues for agreement on the definition and operationalization of an unseen, half-guessed-at construct or phenomenon. The community of scholars has a role to play here. We also refer readers here to Chapter 27 on testing.

14.16 Validity and reliability in life histories

Three central issues underpin the quality of data generated by life history methodology. They concern representativeness, validity and reliability. Plummer (1983) draws attention to a frequent criticism of life history research, namely that its cases are atypical rather than representative. To avoid this charge, he urges researchers to clarify and make explicit the life history in question's relationship to a wider population (Plummer, 1983) by way of appraising the subject on a continuum of representativeness and non-representativeness.

Reliability in life history research hinges on identifying sources of bias and applying techniques to reduce them. Bias arises from the informant, the researcher and the interactional encounter itself. Box 14.1, adapted

BOX 14.1 PRINCIPAL SOURCES OF BIAS IN LIFE HISTORY RESEARCH

Source: Informant
Is misinformation (unintended) given?
Has there been evasion?
Is there evidence of direct lying and deception?
Is a 'front' being presented?
What may the informant 'take for granted' and hence not reveal?
How far is the informant 'pleasing you'?
How much has been forgotten?
How much may be self-deception?

Source: Researcher
Attitudes of researcher: age, gender, class, race, religion, politics etc.
Demeanour of researcher: dress, speech, body language etc.
Personality of researcher: anxiety, need for approval, hostility, warmth etc.
Scientific role of researcher: theory held (etc.), researcher expectancy.

Source: The interaction
The encounter needs to be examined. Is bias coming from:
The physical setting – 'social space'?
The prior interaction?
Non-verbal communication?
Vocal behaviour?

Source: Adapted from Plummer (1983, table 5.2, p. 103)

from Plummer (1983), provides a checklist of some aspects of bias arising from these principal sources.

Several validity checks are available to researchers here. Plummer identifies the following:

1 The subject of the life history may present an autocritique of it, having read the entire product.
2 A comparison may be made with similar written sources by way of identifying points of major divergence or similarity.
3 A comparison may be made with official records by way of imposing accuracy checks on the life history.
4 A comparison may be made by interviewing other informants.

Essentially, the validity of any life history lies in its ability to represent the informant's subjective reality, that is to say, his or her definition of the situation.

14.17 Validity and reliability in case studies

Case studies can use quantitative, qualitative and mixed methods research, and we have indicated earlier in this chapter the canons of validity and reliability for these. Yin (2009) focuses on three main types of validity in case studies: construct, internal and external. Construct validity, he avers, can be addressed by using: (a) multiple sources of evidence (which can lead to convergent lines of enquiry) (pp. 41–2); (b) a chain of evidence (from case study questions to the protocols for the case study – those protocols which link case study questions to the topic in hand, to the evidentiary base, to the conclusions and reporting) (p. 123); and (c) key informants to review drafts of the case study report (p. 41).

Pattern matching can also be used for establishing construct validity (Yin, 2009, p. 41). Here a predicted pattern (a theoretical pattern) in the data is matched to an observed operational pattern and any plausible alternative theories are removed. Pattern matching here can correlate or compare the extent to which what actually occurred reflects the theoretical predictions or explanations (Trochim, 2006).

Internal validity, Yin (2009, p. 41) suggests, can be addressed by: (a) pattern matching; (b) building explanations and considering rival, alternative explanations, and using logic models (where the dependent variable at one stage becomes the independent variable in the next stage of the causal research in a time sequence and in which predicted and observed events are compared (pp. 149ff.)).

Case studies, given their context-specificity and emphasis on subjectivity, may have limited or no external validity; that is, they may not be generalizable. However, external validity, Yin observes, can be addressed by careful use of theory in the single-case studies, such that replication can be conducted (2009, p. 41). He suggests that 'analytic generalizability' (p. 43) may be possible, i.e. where a research strives to generalize from a particular set of findings to some broader or more enduring theory (p. 43). In this he notes the importance of replication studies (p. 44).

External validity here also has to consider the likelihood of transferability of the case study from one context to another, and whether those contexts and the causal connections (Cartwright and Hardie, 2012) between elements in those contexts differ.

Reliability, Yin suggests, benefits from a case study database (2009, p. 41) as this can provide the evidentiary source for checking, together with respondent validation. Yin underlines the importance of keeping careful documentary evidence here (p. 45) and of documenting all aspects and stages of the research so that they can be checked (i.e. transparency).

In essence, the points that Yin makes for addressing validity and reliability in case study research can apply to other types of educational research, and they act as a useful conclusion to this chapter.

Companion Website

The companion website to the book provides additional materials and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Part 3
Methodologies for educational research

It is important to distinguish between design, methodology and instrumentation. Too often methods of data collection are confused with methodology, and methodology is confused with design. Part 2 provided an introduction to design issues and this part examines different methodologies of research, different styles, kinds of, and approaches to, research, separating them from methods – instruments for data collection. We identify key styles of educational research in this section: a bundle of approaches that come under the umbrella of naturalistic/qualitative/ethnographic types of research; historical and documentary research (entirely rewritten by Jane Martin); different kinds of survey, with an entirely new chapter on Internet surveys; case studies; different kinds of experiment, an extended review of randomized controlled trials and coverage of *ex post facto* research; a chapter on meta-analysis, research syntheses and systematic reviews (entirely rewritten by Harsh Suri), which takes account of the increased prominence given to these in the research community; action research; and an entirely rewritten and updated chapter on virtual worlds, social network software and netography in educational research (written by Stewart Martin). These chapters include more extended analysis of key issues and features of the different research styles that researchers can address in planning and implementing their research.

Although we recognize that these are by no means exhaustive, we suggest that they cover the major styles of research methodology. These take in quantitative as well as qualitative research, mixed methods, the emerging field of virtual worlds, social networking and netography, together with small-scale and large-scale approaches. Selecting research approaches is not a matter simply of preference, arbitrary or automatic decision making, but, like other aspects of research, is a deliberative process in which the key is the application of the notion of *fitness for purpose*. We do not advocate slavish adherence to a single methodology in research; indeed combining methodologies may be appropriate for the research in hand. The intention here is to shed light on the different styles of research, locating them in the paradigms of research introduced in Part 1.

The companion website to the book provides additional materials and PowerPoint slides for the chapters in this part, which can be found online at: www.routledge.com/cw/cohen.

Qualitative, naturalistic and ethnographic research

CHAPTER 15

The title of this chapter indicates that a wide range of types and kinds of qualitative research are addressed here. The chapter addresses several key issues in planning and conducting qualitative research:

- foundations of qualitative, naturalistic and ethnographic inquiry (theoretical bases of these kinds of research)
- naturalistic research
- ethnographic research
- critical ethnography
- autoethnography
- virtual ethnography
- phenomenological research
- planning qualitative, naturalistic and ethnographic research
- reflexivity
- doing qualitative research
- some challenges in qualitative, ethnographic and naturalistic approaches

There is no single blueprint for naturalistic, qualitative or ethnographic research, because there is no single picture of the world. Rather, there are many worlds and many ways of investigating them. In this chapter we set out a range of key issues in understanding these worlds.

'Qualitative research' is a loosely defined term that includes a vast range of kinds of research, has a wide range of meanings and covers a heterogeneity of fields (Preissle, 2006; Hammersley, 2013, p. 9), so much so that Hammersley (2013) suggests that, given this range, the term may no longer be a 'genuine or useful category' (p. 99). He quotes Bryman's (2008) view that qualitative research connotes the use of words rather than numbers (p. 366), and Sandelowski's (2001) view that it focuses on the attitudes towards understanding, experiences and interpretations by humans of the social world, and how to enquire about all of these (p. 893) (though, as Hammersley (2013, p. 2) notes, these are not exclusive to qualitative research). Hammersley defines qualitative research as:

> a form of social inquiry that tends to adopt a flexible and data-driven research design, to use relatively unstructured data, to emphasize the essential role of subjectivity in the research process, to study a number of naturally occurring cases in detail, and to use verbal rather than statistical forms of approach.
>
> (Hammersley, 2013, p. 12)

There are several purposes of qualitative research, for example, description, explanation, reporting, creation of key concepts, theory generation and testing. It is important to stress, at the outset, that, though there are many similarities and overlaps between naturalistic/ethnographic and qualitative methods, there are also differences between them. The former connotes long-term residence with an individual, group or specific community (cf. Swain, 2006, p. 206), whilst the latter, often being concerned with the nature of the data and the kinds of research question to be answered, is an approach that need not require naturalistic approaches or principles. That said, there are sufficient areas of commonality to render it appropriate to consider them in the same chapter, and we tease out differences between them where relevant. The intention of this chapter is to provide guidance for qualitative researchers who are conducting either long-term ethnographic research or small-scale, short-term qualitative research.

There are many varieties of qualitative research, indeed Preissle (2006, p. 686) remarks that qualitative researchers cannot agree on its purposes, boundaries, disciplinary fields or, indeed, its terminology (interpretive, naturalistic, qualitative, ethnographic, phenomenological, anthropological, symbolic interactionist, critical theoretical, case study, grounded theory etc.). However, she does indicate that qualitative research is characterized by a 'loosely defined' group of designs that elicit verbal, aural, observational, tactile, gustatory and olfactory information from a range of sources including audio, film, documents and pictures, that it draws strongly on direct experience and meanings, and that

these may vary according to the style of qualitative research undertaken.

Qualitative research provides an in-depth, intricate and detailed understanding of meanings, actions, non-observable as well as observable phenomena, attitudes, intentions and behaviours, and these are well served by naturalistic enquiry (Gonzales *et al*., 2008, p. 3). It gives voices to participants and it probes issues that lie beneath the surface of presenting behaviours and actions.

Qualitative research can be used in systematic reviews (Dixon-Woods *et al*., 2001; see also Chapter 21 of the present volume), to:

- identify and refine questions, fields, foci and topics of the review, i.e. to act as a precursor to a full review;
- provide data in their own right for a research synthesis;
- indicate and identify the outcomes that are of interest, and for whom;
- complement and augment data from quantitative reviews;
- fill gaps in quantitative reviews;
- explain the findings from quantitative reviews and data;
- provide alternative perspectives on topics;
- contribute to the drawing of conclusions from the review;
- be part of a multi-methods research synthesis;
- suggest how to turn evidence into practice.

Whilst the lack of controls in much qualitative research renders it perhaps unattractive for research syntheses, this is possibly unjustified, as it suggests that qualitative research has to abide by the rules of the game of quantitative approaches. Qualitative methods have their own tenets, and these complement very well those of numerical research. As with quantitative studies, qualitative studies have to be weighted, downgraded, upgraded or excluded according to the quality of the evidence and sampling that they contain. They also have to overcome the problem that, by stripping out the context in order to obtain themes and key concepts, they destroy the heart of qualitative research – context (Dixon-Woods *et al*., 2001, p. 131).

15.1 Foundations of qualitative, naturalistic and ethnographic inquiry

The social and educational world is a messy place, full of contradictions, richness, complexity, connectedness, conjunctions and disjunctions. It is multilayered and not easily susceptible to the atomization or aggregation processes inherent in much numerical research. It has to be studied in total rather than in fragments if a true understanding is to be reached. Chapter 1 indicated that several approaches to educational research are contained in the paradigm of qualitative, naturalistic and ethnographic research. The characteristics of that paradigm (Boas, 1943; Blumer, 1969; Lincoln and Guba, 1985; Woods, 1992; LeCompte and Preissle, 1993; Ary *et al*., 2002; Flick, 2009; Larsson, 2009; Hammersley, 2013, 2014; Pring, 2015; Wellington, 2015) can be set out ontologically (what is it we are trying to understand?), epistemologically (how can we know about something?) and methodologically (how can we research something?).

Ontology of qualitative research

- Qualitative research regards people as anticipatory, meaning-making beings who actively construct their own meanings of situations and make sense of their world and act in it through such interpretations (the constructivist/constructionist premise). People are deliberate, intentional and creative in their actions, and meaning arises out of social situations, interactions and negotiations, and is handled through the interpretive processes of the humans involved.
- Meanings used by participants to interpret situations are culture- and context-bound, and there are multiple realities, not single truths in interpreting a situation. History and biography intersect – we create our own futures but not necessarily in situations of our own choosing.
- Realities are multiple, constructed and holistic, capable of sustaining multiple interpretations, including those of all parties involved. People, situations, events and objects are unique and have meaning conferred upon them rather than possessing their own intrinsic meaning. Knower and known are interactive, inseparable.

Epistemology of qualitative research

- Behaviour and, thereby, data are socially situated, context-related, context-dependent and context-rich. To understand a situation researchers need to understand the context both specifically and holistically – the whole picture – because situations affect behaviour and perspectives and vice versa. One task of the researcher is to understand, describe and explain the multiple and differing interpretations of situations, their distinctiveness, causes and consequences.
- All factors, rather than a limited number of variables, have to be taken into account in understanding

a phenomenon. Research looks at relationships between elements in a whole system. As all entities are in a state of mutual simultaneous shaping, it is difficult, if not impossible, or inappropriate to distinguish causes from effects. The attribution of meaning is continuous and evolving over time.

- Social reality, experiences and social phenomena are capable of multiple, sometimes contradictory interpretations and are available to us through social interaction. Researchers focus on subjective accounts, views and interpretations of a phenomenon by the participants (including the researcher): their 'definition of the situation', which is typically reported verbally rather than numerically. Social research examines situations through the eyes of the participants; the task of ethnographies, as Malinowski (1922, p. 25) observed, is to grasp the point of view of the native [*sic*], his [*sic*] view of the world and in relation to his [*sic*] life.

Methodology of qualitative research

- Research must include 'thick descriptions' (Geertz, 1973) of the contextualized behaviour; for descriptions to be 'thick' requires inclusion not only of detailed observational data and data on meanings, participants' interpretations of situations and unobserved factors. Observational data are important, acquired from the natural, undisturbed setting, with participants speaking in their own terms and behaving 'naturally'. To understand and research a situation often requires long-term immersion in the system, not least as researchers do not know in advance what they will see or what they will look for.
- Social research should be conducted in natural, uncontrived, real-world settings with as little intrusiveness as possible by the researcher. Here data are collected systematically, analysed inductively and abductively, with constructs and findings deriving and inferred from the data during the research. Human phenomena seem to require even more conditional stipulations than do other kinds of phenomena, and meanings and understandings replace proof. Only time- and context-bound working hypotheses and idiographic statements are possible, and researchers generate rather than test hypotheses. Theory generation is derivative – grounded (Glaser and Strauss, 1967) – the data suggest the theory rather than vice versa.
- The *processes* of research and behaviour are as important as the outcomes. Research is value-bound and is influenced by the researcher's values as expressed in the choice of the focus of the research, its framing and bounding, method of working and data collection, analysing and reporting findings. Research is influenced by the choice of the paradigm that guides the investigation into the problem, and the choice of the substantive theory utilized to guide the collection and analysis of data and in the interpretation of findings. Research is influenced by the values that inhere in the context, which may be congruent or dissonant within and between the parties involved.
- Researchers are the instruments of the research (Eisner, 1991), blurring the distinction between the researcher and other participants and between subjective and objective facts, as 'objective facts' are mediated through subjective interpretations.
- Generalizability is interpreted as generalizability to identifiable, specific settings and subjects rather than universally. The context-specificity of the phenomenon being research often precludes generalization. Larsson (2009) suggests that generalization can be addressed through maximizing variation, context similarity and recognition of patterns (p. 28).

Hammersley (2013, pp. 29–34) notes the 'critical' tradition in qualitative research, in which situations are observed and interpreted through wide-angle lenses that include a focus on the multiple, intersecting, wider factors that bear on a situation, utilizing ideology critique with an interest in emancipation from oppression, exploitation, inequality, power and powerlessness, and un-freedoms, i.e. research with an overtly political intent to expose the deforming interests (ideologies) at work in a situation and to bring about a more just society.

Lincoln and Guba (1985, pp. 39–43), Ary *et al*. (2002, pp. 451–7) and Polkinghorne (2007) tease out implications of these axioms:

- studies must be set in their natural settings as context is heavily implicated in meaning;
- humans are the research instrument;
- utilization of tacit knowledge is inescapable;
- qualitative methods sit more comfortably than quantitative methods with the notion of the human-as-instrument;
- purposive sampling enables the full scope of issues to be explored;
- data analysis is inductive rather than a priori and deductive;
- theory emerges rather than being pre-ordinate. A priori theory is replaced by grounded theory;
- research designs emerge over time (and as the sampling changes over time);

- the outcomes of the research are negotiated;
- the natural mode of reporting is the case study;
- nomothetic interpretation is replaced by idiographic interpretation;
- applications are tentative and pragmatic;
- the focus of the study determines its boundaries;
- trustworthiness, credibility, theoretical adequacy, corroboration, interpretive adequacy, dependability and confirmability replace more conventional views of reliability and validity. Here a statement or claim for knowledge is not 'intrinsically valid'; rather, validity is a matter of 'intersubjective judgement' as determined by the community of participants (widely defined) and the force of the argument and evidence (Polkinghorne, 2007, pp. 474–5).

Whilst these points above suggest considerations in addressing quality in qualitative research (Hammersley, 2007), researchers will need to note that it is invidious to produce simplistic, single, permanent or universal lists of criteria for quality in qualitative research (Guba and Lincoln, 2005), as they may all too easily be a poor fit to the range of types and methodologies of qualitative research. Tracy (2010, p. 840), whilst being mindful of the dangers in this enterprise (p. 838), sets out and defends 'eight "big tent" criteria for excellent qualitative research':

1. *A worthy topic* (one which is timely, relevant, significant and interesting, pointing out surprises that challenge common-sense assumptions). Pelias (2015) notes that the researcher has to ask whose interests are being served by the research.
2. *'Rich rigor'*, with attention to the theoretical constructs used (matching the complexity of the phenomenon with the complexity of theoretical constructs), contexts, sampling, sufficient data to support the claims made and the levels of analysis applied, data collection and analysis, time in the field and transparency of data and analysis.
3. *Sincerity*: addressing self-reflexivity (discussed below) and introspection, honesty, vulnerability, authenticity and transparency concerning the research process, from entry to the field to exit from it, together with data analysis.
4. *Credibility*: 'trustworthiness, verisimilitude and plausibility of the research findings' (Tracy, 2010, p. 842), addressed by thick description, triangulation, multivocality, autoethnographies, member reflections, crystallization, demonstration of how the findings and conclusions were reached, persuasiveness of the claims in light of the warrants, concrete details and disclosure of tacit knowledge. In this respect Tracy (2010) notes (p. 844) that triangulation and crystallization may sit together uncomfortably, as triangulation suggests a single correct conclusion (accuracy) whilst crystallization (looking through a crystal from many different viewpoints to see many refracted images) yields different findings.
5. *Resonance*: readers find points of resonance with themselves (empathy, identification and reverberation) and are affected by the research through evocative, engaging, vivid and carefully worded representations, transferable findings to their own situation and 'naturalistic generalizations' (applicability to their own situation in improving practice).
6. *Significant contribution*: moving forward the field conceptually, theoretically, methodologically, practically, heuristically, morally and practically, which includes catalytic validity (see Chapter 14).
7. *Ethical*: attention to procedural ethics (see the discussion of ethics later in this chapter), situational (see Chapter 7), cultural, relational (awareness of the researcher's influence on other participants) and the ethics of leaving the research field.
8. *Meaningful coherence*: achievement of the study's purposes, fitness for purposes in methods and procedures used, linkages between literature, research questions, methods, findings and interpretations.

Lincoln and Guba (1985, pp. 226–47) set out ten elements in research design for qualitative studies:

1. Determining a focus for the inquiry;
2. Determining the fit of paradigm to focus;
3. Determining the 'fit' of the inquiry paradigm to the substantive theory selected to guide the inquiry;
4. Determining where and from whom data will be collected;
5. Determining successive phases of the inquiry;
6. Determining instrumentation;
7. Planning data collection and recording modes;
8. Planning data-analysis procedures;
9. Planning the logistics:
    a. prior logistical considerations for the project as a whole
    b. the logistics of field excursions prior to going into the field
    c. the logistics of field excursions while in the field
    d. the logistics of activities following field excursions
    e. the logistics of closure and termination;
10. Planning for trustworthiness.

These elements can be set out into a sequential, staged approach to planning naturalistic research (e.g. Schatzman and Strauss, 1973; Delamont, 1992). Spradley (1979) sets out the stages of: (i) selecting a problem; (ii) collecting cultural data; (iii) analysing cultural data; (iv) formulating ethnographic hypotheses; (v) writing the ethnography. We offer a fuller, twelve-stage model later in this chapter.

Like other styles of research, naturalistic and qualitative methods can formulate research questions which should be clear and unambiguous but open to change as the research develops. Strauss (1987) terms these 'generative questions': they stimulate the line of investigation, suggest initial hypotheses and areas for data collection, yet they do not foreclose the possibility of modification as the research develops. A balance must be struck between having research questions that are so broad that they do not steer the research in any particular direction, and so narrow that they block new avenues of enquiry (Flick, 2004b, p. 150).

Miles and Huberman (1994) identify two types of qualitative research design: loose and tight. Loose research designs have broadly defined concepts and areas of study, and, indeed, are open to changes of methodology. These are suitable, they suggest, when researchers are experienced and when the research is investigating new fields or developing new constructs, akin to the flexibility and openness of theoretical sampling (see Chapter 37). By contrast, a tight research design has narrowly restricted research questions and predetermined procedures, with limited flexibility. These, the authors suggest, are useful when researchers are inexperienced, when the research is intended to look at particular specified issues, constructs, groups or individuals, or when the research brief is explicit.

Even though, in qualitative research, issues and theories emerge from the data, this does not preclude the value of having research questions. Flick (1998, p. 51) suggests three types of research questions in qualitative research, namely, those concerned with: (a) describing states, their causes and how these states are sustained; (b) describing processes of change and consequences of those states; (c) suitability for supporting or not supporting hypotheses and assumptions or for generating new hypotheses and assumptions (the 'generative questions' referred to above).

Should one have a hypothesis in qualitative research?

We mentioned in Chapter 1 that positivist approaches typically test pre-formulated hypotheses and that a distinguishing feature of naturalistic and qualitative approaches is its reluctance to enter the hypothetico-deductive paradigm (e.g. Meinefeld, 2004, p. 153), not least because there is a recognition that the researcher influences the research and because the research is much more open and emergent in qualitative approaches. Indeed, Meinefeld, citing classic studies like Whyte's (1955) *Street Corner Society*, suggests that it is impossible to predetermine hypotheses, whether one wishes to or not, as prior knowledge cannot be presumed. Glaser and Strauss (1967) suggest that researchers should deliberately free themselves from all prior knowledge, even suggesting that it is impossible to read up in advance, as it is not clear what reading will turn out to be relevant – the data speak for themselves. Theory is the end point of the research, not its starting point.

One has to be mindful that the researcher's own background interest, knowledge and biography precede the research and that though initial hypotheses may not be foregrounded in qualitative research, nevertheless the initial establishment of the research presupposes a particular area of interest, i.e. the research and data for focus are not theory-free; knowledge is not theory-free. Indeed Glaser and Strauss (1967) acknowledge that they brought their own prior knowledge to their research on dying.

The resolution of this apparent contradiction – the call to reject an initial hypothesis in qualitative research, yet a recognition that all research commences with some prior knowledge or theory that gives rise to the research, however embryonic – lies in several fields. These include: an openness to data (Meinefeld, 2004, pp. 156–7); a preparedness to modify one's initial presuppositions and position; a declaration of the extent to which the researcher's prior knowledge may be influencing the research (i.e. reflexivity); a recognition of the tentative nature of one's hypothesis; a willingness to use the research to generate a hypothesis; and an acknowledgement that having a hypothesis may be just as much a part of qualitative research as it is of quantitative research.

An alternative to research hypotheses in qualitative research is a set of research questions, and we consider these below. For qualitative research, Miles and Huberman (1994, p. 74) also suggest the replacement of 'hypotheses' with 'propositions', as this indicates that the qualitative research is not necessarily concerned with testing a predetermined hypothesis as such but, nevertheless, is concerned to be able to generate and test a theory (e.g. grounded theory).

15.2 Naturalistic research

Main *kinds* of naturalistic enquiry are (Arsenault and Anderson, 1998, p. 121; Flick, 2004a, 2004b, 2009):

- *case study* (an investigation into a specific instance or phenomenon in its real-life context);
- *comparative studies* (where several cases are compared on the basis of key areas of interest);
- *retrospective studies* (which focus on biographies of participants or which ask participants to look back on events and issues);
- *snapshots* (analyses of particular situations, events or phenomena at a single point in time);
- *longitudinal* studies (which investigate issues or people over time);
- *ethnography* (a portrayal and explanation of social groups and situations in their real-life contexts);
- *grounded theory* (developing theories to explain phenomena, the theories emerging from the data rather than being prefigured or predetermined);
- *biography* (individual or collective);
- *phenomenology* (seeing things as they are really like and establishing the meanings of things through illumination and explanation rather than through taxonomic approaches or abstractions, and developing theories through the dialogic relationships of researcher to researched).

The main *methods* for data collection in naturalistic enquiry are (Hammersley and Atkinson, 1983):

- participant observation
- interviews and conversations
- documents and field notes
- accounts
- notes and memos.

Ary *et al*. (2002) add to these grounded theory and historical research.

Lofland (1971) suggests that naturalistic methods are intended to address three major questions:

- What are the characteristics of a social phenomenon?
- What are the causes of the social phenomenon?
- What are the consequences of the social phenomenon?

These include: (a) the environment; (b) people and their relationships; (c) behaviour, actions and activities; (d) verbal behaviour; (e) psychological stances; (f) histories; and (g) physical objects (Baker, 1994, pp. 241–4).

There are several key differences between the naturalistic approach and that of the positivists to whom we made reference in Chapter 1. LeCompte and Preissle (1993, pp. 39–44) suggest that ethnographic approaches are concerned with description rather than prediction, induction rather than deduction, generation rather than verification of theory, construction rather than enumeration, and subjectivities rather than objective knowledge. With regard to the latter, they distinguish between *emic* approaches (as in the term 'phonemic', where the concern is to catch the subjective meanings placed on situations by participants) and *etic* approaches (as in the term 'phonetic', where the intention is to identify and understand the objective or researcher's meaning and constructions of a situation) (p. 45).

Woods (1992), however, argues that some differences between quantitative and qualitative research are exaggerated, that the epistemological contrast between the two is overstated, as qualitative techniques can be used for both generating and testing theories.

15.3 Ethnographic research

As the 'interview society' is losing ground to the 'observation society' (Gobo, 2011), ethnography, being largely observation-based, is coming into prominence. LeCompte and Preissle (1993) suggest that ethnographic research seeks to create as vivid and analytical a reconstruction as possible of the culture or groups being studied (p. 235). An ethnography is a descriptive, analytical and explanatory study of the culture (and its components), values, beliefs and practices of one or more groups (e.g. Creswell, 2012, p. 462; Bhatti, 2012; Denscombe, 2014). It can study a small group (a few people: micro-ethnography) (p. 463) or a larger group/society/community, and, in autoethnography, an individual in a social setting. Though it typically uses qualitative data, it does not preclude the use of relevant quantitative data (Hamersley, 2006).

LeCompte and Preissle (1993) and Denscombe (2014) indicate several key elements of ethnographic approaches:

- the world view of the participants is investigated and represented – their 'definition of the situation' (Thomas, 1923);
- data are elicited and gathered;
- researchers spend considerable amounts of time in the field – immersion – to research the everyday as well as the non-normal aspects of the culture, group, etc. (though Hammersley (2006) notes that, in comparison to earlier times, much fieldwork includes shorter rather than longer stays in the field, and this may risk loss of important historical contextual information and the danger of assuming that the data

collected are a fair representation of the entire situation);
- meanings are accorded to phenomena by both the researcher and the participants; the process of research, therefore, is hermeneutic, uncovering meanings;
- the constructs of the participants are used to structure the investigation;
- empirical data are gathered in their naturalistic setting (unlike laboratories or in controlled settings as in other forms of research where variables are manipulated);
- observational techniques are used extensively (both participant and non-participant) to acquire data on real-life settings;
- the research is holistic, that is, it seeks a description and interpretation of 'total phenomena';
- there is a move from description and data to inference, explanation, suggestions of causation, and theory generation;
- methods are 'multimodal' and the ethnographer is a 'methodological omnivore' (LeCompte and Preissle, 1993, p. 232).

Hitchcock and Hughes (1989, pp. 52–3) suggest that ethnographies involve:

- the production of descriptive cultural knowledge of a group;
- the description of activities in relation to a particular cultural context from the point of view of the members of that group themselves;
- the production of a list of features constitutive of membership in a group or culture;
- the description and analysis of patterns of social interaction;
- the provision as far as possible of 'insider accounts';
- the development of theory.

Bryman (2008) notes that ethnographic researchers immerse themselves in the group or society which they are studying in order to collect field data which may comprise descriptive notes and analytical comments about the culture of the members of the society or group which they are studying, including the views and definitions of the situation of the members themselves, which are then written up in way that is amenable and accessible to the target audience or readership. An ethnography moves beyond description to data analysis, to theory generation and, if appropriate, to hypothesis generation, to explain what is happening and observed in a situation, group, culture or society and why, what are its key dynamics, in short to understand why the group, culture or society is acting as it does and what can be learned from this.

In educational research, Walford (2009) notes that ethnographies can use multiple data types to focus in depth on cultural formations and maintenance: how the culture works (p. 273). Knowledge about these is obtained through sustained, long-term immersion and involvement in the group, giving importance to the 'accounts of participants' perspectives and understandings' (p. 272), all of which lead to the formation of hypotheses and theory testing which can provide the basis for theoretical generalization and, indeed, subsequent data collection (p. 272). Like Hammersley (2006), Walford does not rule out quantitative data. Ethnography, Walford avers (p. 275), is story-telling, with the researcher centrally involved in the generation and telling of the story (p. 275).

For Denscombe (2014, p. 90), the attractions of ethnographies are that they use detailed and direct observational data; they focus on holism in the research; bring a fresh eye to the obvious, ordinary, taken-for-granted and everyday behaviour; take seriously the participants' views; have strong ecological validity; and are self-aware (see section below, 'Reflexivity'). On the other hand, he notes that ethnographies have to balance objective descriptions of events and cultures with the researcher's own interpretation of these, and they have to avoid the risk of creating isolated 'pictures' of situations which lack an overall structure. Ethnographers and ethnographies must not sacrifice analysis to telling a story. Ethnographies are difficult – if not impossible – to replicate or to check, and they raise difficult ethical issues more prominently than other approaches (discussed below), and researchers may experience difficulties in gaining access to the research setting (p. 91).

Dobbert and Kurth-Schai (1992) urge not only that ethnographic approaches become more systematic but that they study and address regularities in social behaviour and social structure (pp. 94–5). The task of ethnographers (p. 150) is to balance a commitment to catching the diversity, variability, creativity, individuality, uniqueness and spontaneity of social interactions (e.g. by 'thick descriptions'; Geertz, 1973) with a commitment to the task of social science to seek regularities, order and patterns within such diversity. As Durkheim (1982) noted, there are 'social facts'.

Following this line, it is possible to suggest that ethnographic research can address issues of generalizability – a tenet of positivist research – interpreted as 'comparability' and 'translatability' (LeCompte and Preissle, 1993, p. 47). For comparability, the characteristics of the group that is being studied need to be made

explicit so that readers can compare them with other similar or dissimilar groups. For translatability, the analytic categories used in the research as well as the characteristics of the groups are made explicit so that meaningful comparisons can be made to other groups and disciplines.

Spindler and Spindler (1992, pp. 72–4) put forward several hallmarks of effective ethnographies:

- Observations have contextual relevance, both in the immediate setting in which behaviour is observed and in further contexts beyond.
- Hypotheses emerge *in situ* as the study develops in the observed setting.
- Observation is prolonged and often repetitive. Events and series of events are observed more than once to establish reliability in the observational data.
- Inferences from observation and various forms of ethnographic inquiry are used to address insiders' views of reality.
- A major part of the ethnographic task is to elicit socio-cultural knowledge from participants, rendering social behaviour comprehensible.
- Instruments, schedules, codes, agenda for interviews, questionnaires, etc. should be generated *in situ*, and should derive from observation and ethnographic inquiry.
- A transcultural, comparative perspective is usually present, although often it is an unstated assumption, and cultural variation (over space and time) is natural.
- Some socio-cultural knowledge that affects behaviour and communication under study is tacit/implicit, and may not be known even to participants or known ambiguously to others. It follows that one task for an ethnographer is to make explicit to readers what is tacit/implicit to informants.
- The ethnographic interviewer should not frame or predetermine responses by the kinds of questions that are asked, because the informants themselves have the emic, native cultural knowledge.
- In order to collect as much live data as possible, any technical device may be used.
- The ethnographer's presence should be declared and his or her personal, social and interactional position in the situation should be described.

Ethnographic researchers will need to consider whether to employ interviewing at all, as it is a non-natural situation, or, if it is to be used, what form it will take – away from the interviewer as 'miner' (Kvale, 1996) – seeking nuggets of information (see Chapter 25) – and moving towards the interviewer as 'traveller' in a collaborative journey along the road of knowledge creation for both interviewer and interviewee, i.e. as part of a mutually empowering relationship (e.g. Edwards and Holland, 2013, p. 32).

With 'mutual shaping and interaction' between the researcher and participants taking place (Lincoln and Guba, 1985, p. 155), the researcher becomes, as it were, the 'human instrument' in the research (p. 187), building on her tacit knowledge and her propositional knowledge, using methods that sit comfortably with human inquiry, for example, observations, interviews, documentary analysis and 'unobtrusive' methods (p. 187). The advantage of the 'human instrument' is her adaptability, responsiveness, knowledge, ability to handle sensitive matters, ability to see the whole picture and ability to clarify, summarize, explore, analyse and examine atypical or idiosyncratic responses (pp. 193–4). Here Hammersley (1992b) comments on the risk of researcher bias (see section below on 'Reflexivity').

Denscombe (2014) notes that ethnographies can include life histories, and they need to provide thick descriptions and use both idiographic and nomothetic approaches, the former to produce a detailed picture of the unique situation/culture/group and the latter to produce theories that can apply beyond the situation in question.

A key concern for ethnographers is how far out to go in order to understand a situation (macro issues affecting, contextualizing, locating or contributing to the situation in hand) or how far in to go in focusing on a situation (micro ethnography). In other words, if ethnography celebrates holism, what is the whole and how are data about the whole to be gathered (Hammersley, 2006, pp. 6–7)?

15.4 Critical ethnography

One branch of ethnography that resonates with the critical paradigm outlined in Chapter 3 is critical ethnography – 'critical theory in action' (Madison, 2005, p. 13), which, as Thomas (1993, p. vii) suggests, adopts a 'subversive worldview' to conventional traditions of research. Marshall and Rossman (2016) note that critical ethnography has a wide embrace, taking in different kinds of critical theory, queer theory, critical race theory, autoethnography, feminist theories, critical discourse analysis, participatory action research, cultural studies, post-colonial theories and Internet studies.

Whereas conventional ethnography is concerned with what is, critical ethnography concerns itself with what could or should be (Thomas, 1993, p. 4). Here

qualitative, anthropological, participant, observer-based research has its theoretical basis in critical theory (Quantz, 1992, p. 448; Carspecken, 1996; Creswell, 2012). As outlined in Chapter 3, this paradigm is concerned with the exposure of oppression and inequality in society with a view to emancipating individuals and groups towards collective empowerment. In this respect, research is an inherently political enterprise; it is ethnography with a political intent (cf. Thomas, 1993, p. 4). Madison (2005, p. 5) indicates that critical ethnography has an explicit agenda and an 'ethical responsibility' to promote freedom, social justice, equity and well-being. This, he avers, inevitably involves disturbing accepted meanings and disrupting the status quo and purported neutrality of research, together with exposing taken-for-granted, 'domesticated' (Thomas, 1993, p. 7) assumptions that perpetuate the power of the already powerful at the expense of the powerless and the dominated.

Critical ethnography takes power, control, empowerment, privilege, repression, hegemony, victimization, marginalization and social exploitation as problematic and to be changed rather than simply to be interrogated and discovered (Thomas, 1993, p. 6; Creswell, 2012, p. 467). Like ethnography, it catches ethnographic data, but, beyond this, exposes the data to the ideology critique (see Chapter 3).

Like Habermas's *emancipatory interest* (see Chapter 3 of this volume), research is not simply a scientific, technical exercise, nor is it simply a hermeneutic matter of understanding and interpreting a situation; it does not reject these, but it requires the researcher to move beyond them to engage change (Thomas, 1993, p. 19) as a political act, and it must play its part as activism against hegemonic oppression. Here researchers have to consider their own 'positionality' in this enterprise (Madison, 2005, p. 7), i.e. how their research will help to break domination and inequality. Researchers and their research are neither neutral nor innocent. Both subjectivity and objectivity have to be interrogated for their political stances and effects (p. 8) in relation to those being researched (the 'Others') (p. 9); the research has to make a positive difference to the worlds of the 'Others' (the participants). This moves the ethnographer beyond simply being reflexive to being an activist.

This is contentious: on the one hand it suggests that the researcher is an ideologue (rather than, say, a cool theorist); on the other hand, the claim made is that, like it or not, research is a political act, and that this has been hidden in much research.

Carspecken (1996, pp. 4ff.) suggests several key premises of critical ethnography:

- research and thinking are mediated by power relations;
- these power relations are socially and historically located;
- facts and values are inseparable;
- relationships between objects and concepts are fluid and mediated by the social relations of production;
- language is central to perception;
- certain groups in society exert more power than others;
- inequality and oppression are inherent in capitalist relations of production and consumption;
- ideological domination is strongest when oppressed groups see their situation as inevitable, natural or necessary;
- forms of oppression mediate each other and must be considered together (e.g. race, gender, class).

Quantz (1992, pp. 473–4) argues that research is inescapably value-laden in that it serves some interests, and that in critical ethnography researchers must expose these interests and move participants towards emancipation and freedom. The focus and process of research, at heart political, concern issues of power, domination, voice and empowerment (cf. Lather, 1991). In critical ethnography the cultures, groups and individuals being studied are located in contexts of power and interests. These contexts must be exposed, their legitimacy interrogated and the value base of the research itself exposed. Reflexivity is high in critical ethnography. What separates critical ethnography from other forms of ethnography is that in the former, questions of legitimacy, power, values in society and domination and oppression are foregrounded.

How does the critical ethnographer proceed? This is not an easy task, as critical ethnography focuses on, and challenges, taken-for-granted assumptions and meanings, and these may be difficult to expose simply because they are so taken-for-granted, i.e. embedded in our daily lifeworlds and behaviour. In this sense a critical ethnography is untidy, the study emerges rather than being planned in advance; areas of focus emerge as meanings are revealed and challenged from the position of ideology critique (Thomas, 1993, p. 35). It starts with unsettling issues in society and explores them further (Thomas gives the examples of prisons, the social construction of deviance, racism, prejudice and repressive legislation).

Carspecken and Apple (1992, pp. 512–14) and Carspecken (1996, pp. 41–2) identify five stages in critical ethnography (Figure 15.1).

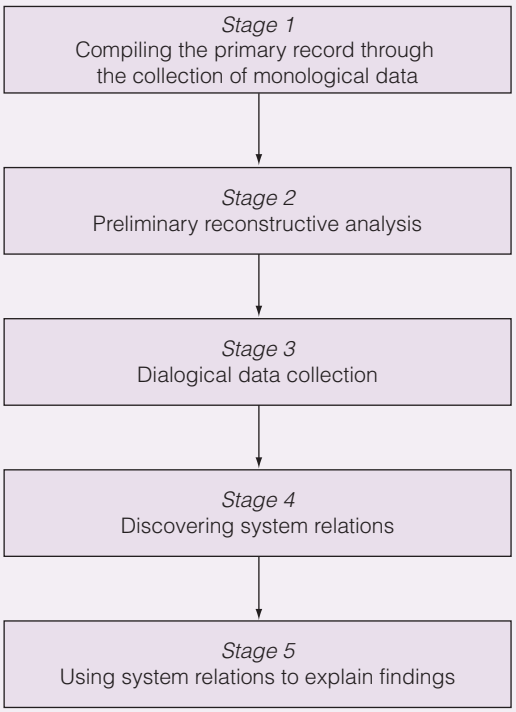


FIGURE 15.1 Five stages in critical ethnography

Stage 1: Compiling the primary record through the collection of monological data

At this stage the researcher is comparatively passive and unobtrusive: a participant observer. The task here is to acquire *objective* data and it is 'monological' in the sense that it concerns only the researcher writing her own notes to herself. Lincoln and Guba (1985) suggest that validity checks at this stage will include:

1. using multiple devices for recording together with multiple observers;
2. using a flexible observation schedule in order to minimize biases;
3. remaining in the situation for a long time in order to overcome the Hawthorne effect;
4. using low-inference terminology and descriptions;
5. using peer-debriefing;
6. using respondent validation.

Echoing Habermas's (1979, 1982, 1984) work on speech-act validity claims, validity here includes truth (the veracity of the utterance), legitimacy (rightness and appropriateness of the speaker), comprehensibility (that the utterance is comprehensible) and sincerity (of the speaker's intentions). Carspecken (1996, pp. 104–5) takes this further in suggesting several categories of reference in objective validity: (i) that the act is comprehensible, socially legitimate and appropriate; (ii) that the actor has a particular identity and particular intentions or feelings when the action takes place; (iii) that objective, contextual factors are acknowledged.

Stage 2: Preliminary reconstructive analysis

Reconstructive analysis attempts to uncover the taken-for-granted components of meaning or abstractions that participants have of a situation. Such analysis is intended to identify the value systems, norms and key concepts that are guiding and underpinning situations. Carspecken (1996, p. 42) suggests that the researcher go back over the primary record from stage one to examine patterns of interaction, power relations, roles, sequences of events, and meanings accorded to situations. He asserts that what distinguishes this stage as 'reconstructive' is that cultural themes and social and system factors that are not usually articulated by the participants themselves are, in fact, reconstructed and articulated, turning the undiscursive into discourse. Moving to higher-level abstractions, this stage can utilize high-level coding (see the discussion of coding below).

In critical ethnography, Carspecken (p. 141) recommends several ways to ensure validity at this stage:

1. use interviews and group discussions with the subjects themselves;
2. conduct member checks on the reconstruction in order to equalize power relations;
3. use peer debriefing (a peer is asked to review the data to suggest if the researcher is being too selective, e.g. of individuals, of data, of inference) to check biases or absences in reconstructions;
4. employ prolonged engagement to heighten the researcher's capacity to assume the insider's perspective;
5. use 'strip analysis' – checking themes and segments of extracted data with the primary data, for consistency;
6. use negative case analysis.

Stage 3: Dialogical data collection

Here data are generated by, and discussed with, the participants (Carspecken and Apple, 1992). The authors argue that this is non-naturalistic in that the participants are being asked to reflect on their own situations, circumstances and lives and to begin to theorize about

their lives. This is a crucial stage because it enables the participants to have a voice, to democratize the research. It may be that this stage produces new data which challenge the preceding two stages.

In introducing greater subjectivity by participants into the research at this stage, Carspecken (1996, pp. 164–5) proffers several validity checks, for example: (a) consistency checks on interviews that have been recorded; (b) repeated interviews with participants; (c) matching observation with what participants say is happening or has happened; (d) avoiding leading questions at interview, reinforced by having peer debriefers check on this; (e) respondent validation; (f) asking participants to use their own terms in describing naturalistic contexts, and encouraging them to explain these terms.

Stage 4: Discovering system relations

This stage relates the group being studied to other factors that impinge on that group, for example: local community groups, local sites that produce cultural products. At this stage Carspecken (1996, p. 202) notes that validity checks will include: (i) maintaining the validity requirements of the earlier stages; (ii) seeking a match between the researcher's analysis and the commentaries that are provided by the participants and other researchers; (iii) using peer debriefers and respondent validation.

Stage 5: Using system relations to explain findings

This stage seeks to examine and explain the findings in light of macro-social theories (Carspecken, 1996, p. 202). In part, this is a matching exercise to fit the research findings within a social theory.

In critical ethnography, therefore, the move is from describing a situation to understanding it, to questioning it and to changing it. This parallels the stages of ideology critique set out in Chapter 3:

Stage 1: a description of the existing situation – a hermeneutic exercise;
Stage 2: a penetration of the reasons that brought the situation to the form that it takes;
Stage 3: an agenda for altering the situation;
Stage 4: an evaluation of the achievement of the new situation.

Critical ethnographies can also be conducted online (Evans, 2010), and we turn to this later in the present chapter.

15.5 Autoethnography

Autoethnography, a derivative of ethnography, is a process, method and product that 'seeks to describe and systematically analyze (*graphy*) personal experience (*auto*) in order to understand cultural experience (*ethno*)' (Ellis *et al*., 2011, p. 1; cf. Reed-Danahay, 1997) and to 'extend sociological understanding' (Wall, 2008, p. 39) by looking at oneself in a wider context. Autoethnographies are 'highly personalized accounts that draw upon the experience of the author/researcher for the purposes of extending sociological understanding' (Sparkes, 2000, p. 21). For examples of this, see Reed-Danahay (1997), Ellis (2004) and Chang (2008).

An autoethnography places the self – the researcher – at the centre of research about himself/herself in a social context; it is self-focused (Ngunjiri *et al*., 2010), though it can engage collaborative as well as individual study (Denshire, 2014). It examines the relationship of the researcher to others through the eyes of the researcher and connects the personal, autobiographic to the social and cultural (Ellis, 2004, p. xix).

Autoethnography often has a deliberate political, social, critical theoretical and emancipatory or transformative agenda (Belbase *et al*., 2008; Bettez, 2015; see Chapter 3 and the comments above on critical ethnography) and it focuses on 'things that matter a great deal to the autoethnographer' (Delamont, 2009, p. 57). It problematizes and interrogates the socially constructed self and the situatedness and relationship of self to others (Starr, 2010, p. 3). For educationists, it has been likened to a form of critical pedagogy in its commitment to transformative and emancipatory processes and the social construction of knowledge (Starr, 2010, p. 4).

Autoethnography concerns studying and writing about our personal and socio-cultural selves, identities and the human condition (Dyson, 2007; Nicol, 2013; Marshall and Rossman, 2016, p. 24), on the assumption that the individual mirrors a social group (Walford, 2009, p. 276). Autoethnography differs from autobiography in that in the latter the focus is only on the self, whilst in the former the focus is on the self in context, typically a socio-cultural context. Denshire (2014) notes that autoethnography moves beyond autobiography 'whenever writers critique the depersonalizing tendencies that can come into play in social and cultural spaces that have asymmetrical relations of power' (p. 833). Autoethnography here concerns how one is 'othered' (Hamilton *et al*., 2008, p. 22) and how one's 'positionality' (discussed later in this chapter) affects the researcher and what is researched (Starr, 2010).

Here emphasis is placed on the writing of the research in a personal, authentic, vivid, engaging and

evocative style, 'writing from the heart' (Denzin, 2006, p. 422) and celebrating the researcher's 'voice' (Wall, 2006, p. 3). As Ellis and Bochner (2006) note, autoethnography catches passion, feelings, struggles, i.e. to evoke the empathy, emotions and sympathy of the reader, indeed for the reader to take action (p. 433), with ideas grounded in the personal experiences of their author, and written in a way that promotes empathy between readers and the 'other' in research.

By contrast, Anderson (2006) and Atkinson (2006) argue for an 'analytic' rather than 'evocative' stance to doing and writing autoethnography, and Anderson (2006) sets out five features of analytic autoethnography: complete member research status (the researcher is a member of the social world being studied); analytic reflexivity (an awareness of, and introspection about, the reciprocal influence of settings, data and researcher) (p. 382); narrative visibility of the researcher's self; dialogue with informants beyond the self; and commitment to theoretical analysis.

Autoethnography recognizes the unavoidable influence of the researcher on the research process, and raises reflexivity (discussed below), subjectivity, emotionality, personal characteristics of the researcher and autobiography to new prominence in the research (cf. Wall, 2006, 2008; Nicol, 2013). It focuses on, and reflects on, the views, 'confessional tales' (Van Maanen, 1988) and analyses of the author on the personal experiences of self and others included in his or her experiences. The author is the participant, looking at himself/herself in socio-cultural locations and terms. In implicating others (often family members, friends and social contacts) in that personal ethnography, ethical issues are raised concerning the confidentiality, anonymity, privacy, safety and protection, non-identifiability and non-traceability of those others and their communities, and relations (sometimes intimate) between the researcher and his/her circle of contacts, social circles and workplace groups. Not only are the 'others' vulnerable, but so is the researcher, the subject of the autoethnography, as self-disclosure about sensitive personal issues can be damaging (Ngunjiri *et al*., 2010).

The consequences of the written autoethnography for the author and others included have to be considered (e.g. Wall, 2006, 2008). This may lead to the need for respondent validation (raising issues of what happens if the respondent wishes to veto data) (Bettez, 2015), or for the removal of identifying features, or changing identifying features (e.g. gender, race, location, appearance) (Ellis *et al*., 2011). The researcher also has to consider the danger of selective memory on his/her part, for example, we may recall but unconsciously distort vivid experiences, raising issues of the credibility and trustworthiness of the report, and these are ethical matters in themselves (cf. Wall, 2006, 2008; Bochner, 2007; Walford, 2009).

Further, in seeking an expressive, evocative style of writing with the intent of reconstructing the authenticity of a lived experience and persuading, touching or moving the reader, there is the danger of subordination of the facts of the case to the emotional response; whether this is legitimate is a moot point – it may be acceptable or out of court. Anderson (2006), for example, as mentioned above, argues for a more analytical than evocative approach to writing autoethnography, whilst Denshire (2014) argues that autoethnography is an essential part of a 'fictive tradition' (p. 836).

In terms of method, autoethnography combines autobiography with ethnography (Ellis *et al*., 2011, p. 2), as the researcher reviews personal experience reflexively, usually retrospectively, and from this analyses and distils key issues about that autobiography from an ethnographic stance, i.e. what the personal experiences say to the reader about culture, values, relations and society in relation to the topic of research interest. This may include writing about moments of existential crisis, turning points ('epiphanies') and life-changing moments.

Autoethnography uses the common tools of ethnography, such as field notes, documents, self-observation and observation of others, interviews, dialogues and conversations (though see the comments later in this chapter about interviews), thick descriptions, reflexivity, grounded theory, long-term attachment and observations of events, times, locations, personal accoutrements such as clothing and artefacts, relationships, power and social life.

An autoethnography is often written in the first person and uses emotional terms, in contrast to much standard academic writing which deliberately adopts a third-person, passive voice and neutral, objective tone. It is often written as a story or narrative and as a personal experience (Marshall and Rossman, 2016, p. 24), with an aesthetic sense as well as a factual basis (Ellis *et al*., 2011). Such storied texts may focus on inequality, oppression, exploitation, subordination, lack of understanding or acceptance (e.g. issues of gender, sexuality, race), injustice, marginalization, stigmatization and 'dominant discourses' (Ellis and Bochner, 2000; Wall, 2008). Less politically or critically, autoethnography may concern issues or experiences that are important to the researchers (e.g. Dyson, 2007; Nicol, 2013). Writing an autoethnography can be therapeutic and cathartic as well as constituting a method of enquiry (Richardson, 2000; Roth, 2009).

Autoethnography is not without its critics (Anderson, 2006; Delamont, 2007, 2009; Ellis *et al*., 2011). For example, it is accused of being an indulgence of the writer, conflating autobiography with research, lacking analytical and theoretical rigour, failing to generate or to test hypotheses or theories, bringing emotions into what should be neutral work, and making reflexivity a thing in itself rather than a tool of ethnography (e.g. Atkinson, 1997; Sparkes, 2000; Wall, 2006). It might be good for the writer, privileging the self (Hamilton *et al*., 2008, p. 17), but of little use to others.

Delamont (2007) argues trenchantly that 'autoethnography is essentially lazy, literally lazy and also intellectually lazy' (p. 1), that it cannot fight familiarity, that it violates ethical standards of privacy and permissions for identifying individuals in published research, that it sacrifices analytical outcomes to reporting of experience, that it focuses on the 'wrong side of the power divide' (p. 5), that it abrogates the duty of the social scientist to go out and gather data rather than 'sit in our offices obsessing about ourselves' (p. 5), that we are not sufficiently interesting to warrant attention by others, and that it is antithetical to two tenets of social science, which are to study the social world and to move the discipline forward. In short, as she writes (Delamont, 2009), it is an 'intellectual *cul de sac*' (p. 51).

Critics contend that it lacks genuine fieldwork and is the apotheosis of navel-gazing, narcissism and self-absorption (e.g. Atkinson, 2006; Madison, 2006; Delamont, 2007, 2009; Ellis *et al*., 2011), i.e. it is more about the 'auto' than the 'ethnography' (Atkinson, 2006, p. 402; Roth, 2009, p. 5). It stands accused of an absence of social context, social action and interaction, of not being sufficiently social to qualify as social science, and of operating in a social vacuum (Atkinson, 1997, p. 339). Delamont (2009), noting that ethnographic research is demanding and hard, derides autoethnography as 'an abrogation of the honourable trade of the scholar' (p. 61).

Such criticisms have been roundly refuted, arguing that differences of views of research should be celebrated rather than proscribed (e.g. Ellis and Bochner, 2000; Bochner, 2001; Denzin, 2006; Wall, 2006, 2008; Starr, 2010; Ellis *et al*., 2011; Denshire, 2014).

15.6 Virtual ethnography

As the Internet is a means of searching a repository of knowledge, a means of communication and a venue for connecting people – real or virtual (Marshall and Rossman, 2016, p. 30) – so the cyberworld has, itself, become a source of ethnographic research. Online communication is a routine and integral part of people's everyday lives, and, given this, its part in ethnographic research is unsurprising. Researchers can enter the Internet and study what is happening in and through it with respect to cultures and communities; the Internet is 'a place where cultural and social phenomena happen' (Webster and da Silva, 2013, p. 123) and where ethnographic interviews can be conducted online (Hanna, 2012).

The Internet is a 'socially constructed space' (Marshall and Rossman, 2016, p. 30), albeit a virtual space (Hine, 2000, 2004), peopled by interacting participants with real and virtual lives, their own cultures, online communities, groups, rites of passage, negotiated roles, group membership and behaviours, and these can be researched as one would conduct an ethnography. The virtual, online environment is the site for the research (Evans, 2010), requiring different, computer-based tools for conducting the research. The computer screen is the on-screen location of the research, and the majority of the data is likely to be text-based, though this does not preclude other data types which are increasingly available on the Internet, for example, Skype, Blackboard Collaborate (Webster and da Silva, 2013).

'Virtual ethnography' (Hine, 2004), netnography (Kozinets, 2010), netography and 'webnography' (Evans, 2010) can be conducted though social networking media, email, online interviews, message boards and messaging, bulletin boards, blogs, chat rooms, forums, discussion boards (see Chapters 23 and 25). The researcher, as in traditional ethnography, is still a participant observer or non-participant observer (Evans, 2010), permanently or intermittently immersed in and observing the virtual environment and what is happening in it, keeping systematic field notes (Hine, 2000).

Because virtual ethnography works with virtual people and alter egos (e.g. avatars), the researcher is often deprived of assurances of honesty and of several features of face-to-face ethnography conducted in the physical presence of the ethnographer and the 'real' participants in their real, physical, 'natural habitat' (Hallett and Barber, 2014, p. 306) (e.g. knowledge of gender, race, age, social status) (Hammersley, 2006, p. 8); it works *as if* participants are real – which they may or may not be – and the 'natural habitat' is now the 'online habitat' (Hallett and Barber, 2014, p. 308).

Netographies overcome problems of time, location and space; they enable the anonymity, privacy and security of the real people to be respected, though this renders problematic the issues of identity and authenticity of the world being investigated. In short, the virtual world is a projection, true or false, of the face-to-face world; it may be no more or less real because of

this (Boellstorff, 2015). The ethnographer proceeds *as if* the Internet world is the real thing, working with the data provided on the Internet, with few, if any, checks on the correctness or authenticity of the actual people behind the avatars. As with other forms of online research, virtual ethnographies raise ethical issues of confidentiality, privacy, anonymity, disclosure, protection from harm to self and others, and informed consent (see also Chapter 8).

For educationists, virtual ethnography can focus on 'real people' in their virtual communities (Hallett and Barber, 2014, p. 310), and on the data which they provide online rather than focusing on virtual people or avatars. This questions how far these are real, full-blooded ethnographies or simply partial and selective data posted online about specific topics of communal or shared interest by interested parties, i.e. extended discussions or sharing of opinions. Indeed Evans (2010) question whether a virtual ethnography is, in fact, more like an extended online survey than an ethnography defined as a 'faithful reproduction of a particular cultural setting' (p. 7).

In conducting a virtual ethnography, Evans (2010) suggests that researchers identify relevant 'proactive communities' (p. 9), raising issues of access, gatekeepers and ethical issues of privacy, anonymity, informed consent, covert or overt research, and permissions. Then researchers can identify key informants and key participants, negotiating access to people and groups and addressing the same ethical issues, with informed consent including both the process and product of the ethnography, and the audience and dissemination of the ethnography. Kulavuz-Onal and Vásquez (2013) remind researchers that they may need to register as a member of an online community in order to gain access and may need to be an active participant in some events whilst being able to be less active in others (p. 229). After this, the researcher can make further contact in order to commence the research, engaging in interaction with the participants (if participant observation is selected) or being a non-participant observer (though Kozinets (2010) advocates participant observation). The researcher gathers ongoing systematically collected and systematically reviewed data and field notes (Kulavuz-Onal and Vásquez, 2013); indeed, online, digital data (including online interviews, see Chapter 25) may lend themselves to software for data analysis. Kulavuz-Onal and Vásquez (2013) comment that fieldwork practices in ethnographies of online communities need to be 'reconceptualised' (p. 237) because they differ from practices in 'in-person ethnographic fieldwork', being software based and computer mediated.

Then the researcher will need to write the ethnography and the report, and seek respondent validation and member checks. This sequence echoes Kozinets's (2010) comments that the methods of traditional ethnographers – gaining entry to the field and community, collecting data, careful analysis and interpretation of data, and reporting, all couched within ethical principles – apply to netographers.

Whilst online research catches some of the social space and topical issues in the community of participants, whether this is sufficient to be called a true, fully fledged, genuine ethnography in the traditional sense of catching the all-round, overall, holistic picture of participants and their socio-cultural settings, is an open question. They are communities united by, or formed by, a common interest rather than having any other connections.

Whilst traditional ethnography sees participants in many settings, presenting many faces and aspects of self to many parties, and whilst participants may present different faces in virtual ethnographies, whether this happens sufficiently in a virtual ethnography for it to be counted as a full-blooded ethnography (rather than, for example, differing views on a topic or different stories from participants) is another open question.

Webster and da Silva (2013) and Hallett and Barber (2014) suggest that, in reality, to conduct a full ethnography could require researchers to study the same participants both online and offline, as the online world is as much part of their 'real' daily lives as the offline, face-to-face, physical interactions. It is a false dualism to separate the online and offline worlds of participants.

15.7 Phenomenological research

Phenomenological research is based on the view that our knowledge of the world is rooted in our (immediate) experiences, and the task of the researcher is to describe, understand, interpret and explain these experiences (Hammersley, 2013, p. 27; Denscombe, 2014, pp. 94–5). This type of research aims to describe, explain and interpret a phenomenon, situation or experience by identifying the meaning of these as understood by the participants, often at an individual as well as a group level (Marshall and Rossman, 2016, pp. 16–17).

As there are many participants involved, each of whom has his/her own authentic meaning and interpretation, there will be multiple realities and accounts; the researcher has to put to one side any prior concepts or suppositions (pp. 27–8) and seek to understand how everyday events and 'commonsense knowledge' (p. 28) are as they are, how they are perceived and

sustained by the participants, and what are the attitudes of participants towards them. In this enterprise, emphasis is placed on the fully described subjective experiences, perceptions, interpretations, attitudes, beliefs, values, feelings and meanings of agentic individuals (Denscombe, 2014, p. 94). In full depiction of lived experiences through the eyes of participants, in come rich description and fidelity to the original experience and out go categorization, abstraction, over-interpretation, quantification and even theorization (pp. 95–6).

Ary *et al*. (2002, p. 447) note that, whilst this is common to much qualitative research, the distinctive feature of phenomenological research is its focus on the subjective experiences of the participants, which are at the heart of the research; what they mean for the participants rather than, for example, the objective 'status of experiences' (p. 447). Not only is there an individual construction of reality, but a social construction of reality (Berger and Luckman, 1967), i.e. there is a shared rather than a solipsistic understanding of the 'real', with shared and multiple realities.

To understand the meanings that participants give to the experiences typically requires in-depth, open-ended and often unstructured interviews with the participants (Marshall and Rossman, 2016, p. 18), which seek to grasp the essence of the meaning(s) of a situation as given by each participant, with detailed descriptions figuring highly here and a recognition that complexity rather than unity or simplicity may be the hallmarks of the meanings given. The researcher has to strive to set aside any of his/her own values, beliefs, taken-for-granted conceptual frameworks, predispositions and everyday background and to see the experience for what it is in the eyes of the participants, freed from such researcher preconceptions, in other words to act as a 'stranger' (Denscombe, 2014, p. 99).

Denscombe writes that phenomenology is suited to small-scale research, descriptive detail of authentic experiences and sympathy to humanistic research which focuses on 'lived experiences'. On the other hand, he notes that, in its pursuit of rich, individualized descriptions, phenomenological research may lack the scientific tenets of 'objectivity, analysis and measurement' (p. 103), may not move beyond description (e.g. to analysis and explanation), may not be generalizable and may focus on trivial everyday events to the neglect of bigger issues (p. 103).

15.8 Planning qualitative, naturalistic and ethnographic research

In many ways the planning issues in qualitative research are not exclusive; they apply to other forms of research, for example: identifying the problem and research purposes; deciding the focus of the study; identifying the research questions; selecting the research design and instrumentation; addressing validity and reliability; ethical issues; approaching data analysis and interpretation. These are common to all research. More specifically, Wolcott (1992, p. 19) suggests that naturalistic researchers should address the stages of watching, asking and reviewing, or, as he puts it, experiencing, enquiring and examining. It is possible to formulate stages in planning naturalistic research (Hitchcock and Hughes, 1989, pp. 57–71; Bogdan and Biklen, 1992; LeCompte and Preissle, 1993). These are presented in Figure 15.2 and are subsequently dealt with in the later pages of this chapter.

One has to be cautious here: Figure 15.2 suggests a linearity in the sequence; in fact, the process is often more complex that this, with a backwards-and-forwards movement between the several stages over the course of the planning and conduct of the research. The process is iterative and recursive, as different elements come into focus and interact with each other in different ways at different times. Indeed Flick (2009, p. 133) suggests a circularity or mutually informing nature of elements of a qualitative research design. In this instance the stages of Figure 15.2 might be better presented as interactive elements as in Figure 15.3.

Further, in some smaller-scale qualitative research not all of these stages may apply, as the researcher may not always be staying for a long time in the field but might only be gathering qualitative data on a 'one-shot' basis (e.g. a qualitative survey, qualitative interviews). However, for several kinds of naturalistic and ethnographic study in which the researcher intends to remain in the field for some time, the several stages set out here, and commented upon in the following pages, may apply.

These stages are shot through with a range of issues that affect the research, for example:

- personal issues (the disciplinary sympathies of the researcher, researcher subjectivities and characteristics, personal motives and goals of the researcher). Hitchcock and Hughes (1989, p. 56) indicate that there are serious strains in conducting fieldwork because the researcher's own emotions, attitudes, beliefs, values, characteristics enter the research; indeed, the more this

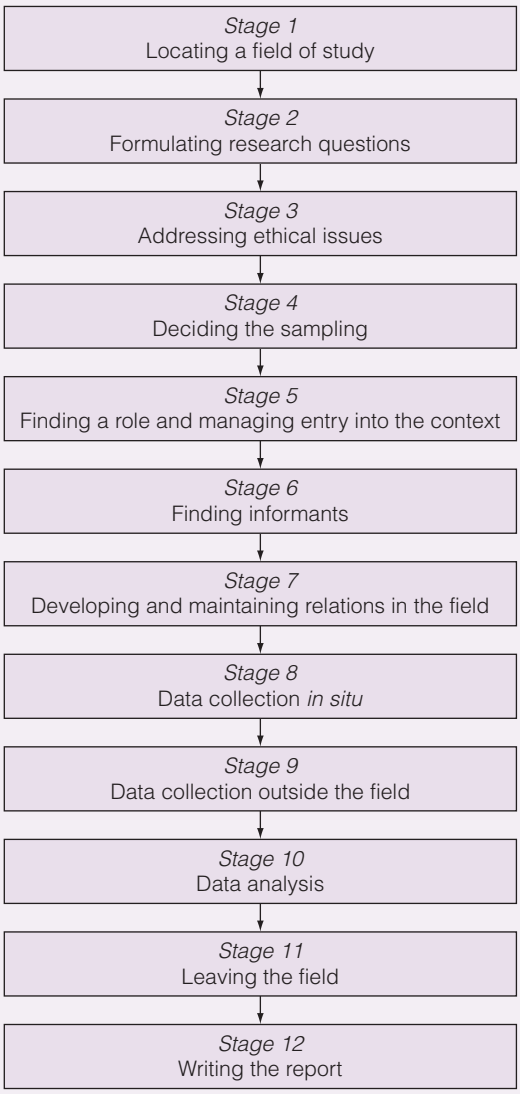


FIGURE 15.2 Stages in the planning of naturalistic, qualitative and ethnographic research

happens the less will be the likelihood of gaining the participants' perspectives and meanings;

- the kinds of participation that the researcher will undertake;
- issues of advocacy (where the researcher may be expected to identify with the same emotions, concerns and crises as the members of the group being studied and wishes to advance their cause, often a feature that arises at the beginning and the end of the research when the researcher is considered to be a legitimate spokesperson for the group);
- role relationships;
- boundary maintenance in the research;
- the maintenance of the balance between distance and involvement;
- ethical issues;
- reflexivity.

15.9 Reflexivity

Reflexivity is a central component of, and a 'crucial strategy' in, qualitative research (Berger, 2015). Researchers have a central role in the creation of knowledge in qualitative enquiry, hence they need to look at themselves and their 'positionality' (discussed later) as part of the research process. Reflexivity recognizes that researchers are inescapably part of the social world that they are researching (Hammersley and Atkinson, 1983, p. 14; Atkinson, 2006, p. 402), and, indeed, that this social world is one already interpreted by the actors, undermining the notion of objective reality. Researchers are in the world and of the world that they research. They bring their own biographies and values to the research situation and participants behave in particular ways in their presence. As Denscombe (2014, p. 88) notes, the researcher does not commence the research 'with a clean sheet', but uses conceptual tools which derive from several sources, including culture and values. What we focus on, what we see, how we understand, describe, interpret and explain are shaped by ourselves and what we bring to the situation. We cannot stand outside these.

Qualitative inquiry is not a neutral activity, and researchers are not neutral; they have their own values, biases and world views, and these are lenses through which they look at and interpret the already-interpreted world of participants (cf. Preissle, 2006, p. 691). Researcher bias is a key issue in qualitative research (as it is with quantitative research) (Hammersley, 1992a).

Researchers, then, have to self-appraise their role in the research process and product (Berger, 2015, p. 220). Pillow (2010) and Bettez (2015) note that reflexive researchers bring their own personal characteristics, experiences, knowledge, backgrounds, values, beliefs, theories, age, gender, sexuality, politics, theories, race, ethnicity, conceptual frameworks and prejudices to the research and that these are often mediated through, and are in conjunction with, issues of power and status. They influence every stage of the research and affect the rapport and conduct of the research. They can affect the formulation of the research topic and questions, access to the field, relationships with participants, data collection, analysis and interpretation, insider and

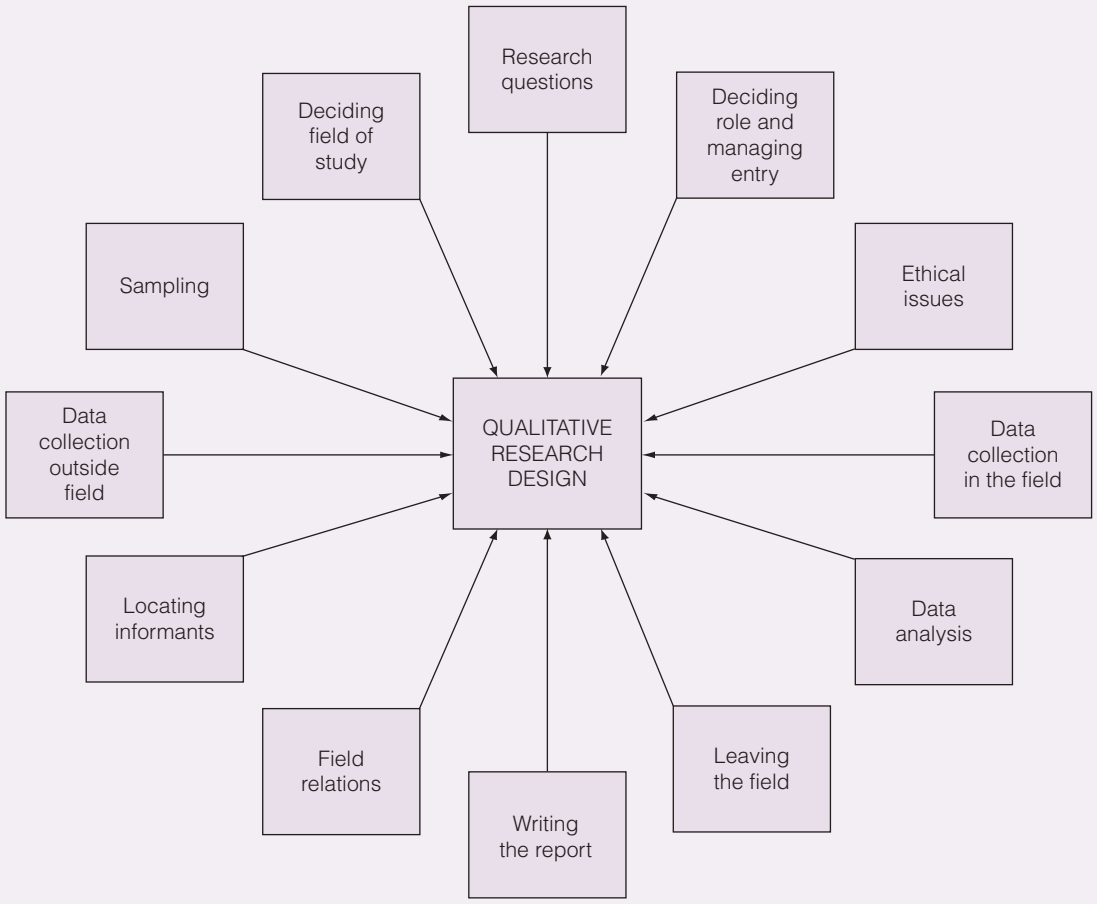


FIGURE 15.3 Elements of a qualitative research design

outsider research, and so on. In short, the researcher may project something or a lot about themselves onto the research (Berger, 2015).

Reflexivity suggests that researchers should consciously and deliberately acknowledge, interrogate and disclose their own selves in the research, seeking to understand their part in, and influence on, the research. Rather than trying to eliminate researcher effects (which is impossible, as researchers are part of the world that they are investigating), researchers should hold themselves up to the light, echoing Cooley's (1902) notion of the 'looking glass self', and researchers should go beyond private reflection on how their own biographies and backgrounds have influenced the research and disclose this publicly as part of the necessary transparency of the research.

Highly reflexive researchers will be acutely aware of the ways in which their selectivity, perception, background, values and inductive processes, frames and paradigms shape the research. They are research instruments. McCormick and James (1988, p. 191) argue that combating reactivity through reflexivity requires researchers to monitor closely and continually their own interactions with participants, their own reactions, roles and biases, and any other matters that might affect the research. This is addressed more fully in Chapter 14 on validity, encompassing issues of triangulation and respondent validity.

15.10 Doing qualitative research

An effective qualitative study has several features (Creswell, 1998, pp. 20–2), and these can be addressed in evaluating qualitative research:

- it uses rigorous procedures and multiple methods for data collection;

- the study is framed within the assumptions and nature of qualitative research;
- enquiry is a major feature, and can follow one or more different traditions (e.g. biography, ethnography, phenomenology, case study, grounded theory);
- the project commences with a focus on an issue, a group, a problem rather than having a hypothesis or the supposition of a causal relationship of variables; relationships may emerge later, but that is open;
- criteria for verification are set out, together with rigour in writing the report;
- verisimilitude is required, such that readers can imagine being in the situation;
- data are analysed at different levels; they are multilayered;
- the writing engages the reader and is replete with unexpected insights, whilst maintaining believability and accuracy.

Maxwell (2005, p. 21) argues that qualitative research should have both *practical* goals (e.g. that can be accomplished, that deliver a specific outcome and meet a need) and *intellectual* goals (e.g. to understand or explain something). His practical goals (p. 24) are: (a) to generate 'results and theories' that are credible and comprehensible to participants and other readers; (b) to conduct formative evaluation in order to improve practice; and (c) to engage in 'collaborative and action research' with different parties. His intellectual goals (pp. 22–3) are: (a) to understand the meanings attributed to events and situations by participants; (b) to understand particular contexts in which participants are located; (c) to identify unanticipated events, situations and phenomena and to generate grounded theories that incorporate these; (d) to understand processes that contribute to situations, events and actions; and (e) to develop causal explanations of phenomena.

He suggests that, whilst quantitative research is interested in discovering the variance and regularity in the effects of one or more particular independent variables on an outcome, qualitative research is interested in the causal processes at work in understanding how one or more interventions or factors lead to an outcome, the mechanisms of their causal linkages. Quantitative research can tell us correlations, how much, whether and 'what', whilst qualitative research can tell us the 'how' and 'why' – the processes involved in understanding and explaining how things occur.

Maxwell also argues that qualitative research should be based on a suitable theoretical basis or paradigm. Quoting Becker (1986), he argues that if a researcher bases his or her research in an inappropriate theory or paradigm it is akin to a worker wearing the wrong clothes: it inhibits comfort and the ability to work properly. Maxwell notes that theoretical premises may not always be clear at the outset of the research; they may emerge, change, be added to etc. over time as the qualitative research progresses (see Chapters 1 and 4 on paradigms and theories). Theory, Maxwell avers (p. 43), can provide a supporting set of principles, world view or sense-making referent, and it can be used as a 'spotlight', illuminating something very specific in a particular event or phenomenon. He advocates a cautious approach to the use of theory (p. 46), steering between, on the one hand, having it unnecessarily constrain and narrow a field of investigation and being accepted too readily and uncritically, and, on the other hand, not using it enough to ground rigorous research. Theories in qualitative research should be those of the researcher and the participants. He suggests that theory can provide the conceptual and justificatory basis for the qualitative research being undertaken, and it can also inform the methods and data sources for the study (p. 55).

Against this background, we set out a twelve-stage process for doing qualitative research.

Stage 1: locating a field of study

Bogdan and Biklen (1992, p. 2) suggest that research questions in qualitative research are not framed by simply operationalizing variables as in the positivist paradigm. Rather, they are formulated *in situ* and in response to situations observed, i.e. topics are investigated in all their complexity in the naturalistic context.

In some qualitative studies, the selection of the research field will be informed by the research purposes, the need for the research, what gave rise to the research, the problem to be addressed and the research questions and sub-questions. In other qualitative studies these elements may only emerge after the researcher has been immersed for some time in the research site itself.

Stage 2: formulating research questions

Research questions are an integral and driving feature of qualitative research. They must be able to be answered concretely, specifically and with evidence (see Chapter 10). They must be achievable and finite (cf. Maxwell, 2005, pp. 65–78) and are often characterized by being closed rather than open questions. Whereas research purposes can be open and less finite, motivated by a concern for 'understanding', research questions, by contrast, though they are informed by research purposes, are practical and able to be accomplished (Maxwell, 2005, pp. 68–9).

Hence, instead of asking a non-directly answerable question such as 'how should we improve online

learning for biology students?', we can ask a specific, focused, bounded and answerable question such as 'how has the introduction of an online teacher–student chat room improved Form 5 students' interest in learning biology?'. Here the word 'should' (as an open question) has been replaced with 'has', the general terminology of 'online learning' has been replaced with 'an online teacher–student chat room' and the open-endedness of the first question has been replaced with the closed nature of the second (cf. Maxwell, 2005, p. 21).

Whereas in quantitative research, a typical research question asks 'what' and 'how much' (e.g. 'how much do male secondary students prefer female teachers of mathematics, and what is the relative weighting of the factors that account for their preferences?'), a qualitative research question often asks more probing, process-driven research questions (e.g. 'how do secondary school students in school X decide their preferences for male or female teachers of mathematics?').

Maxwell (2005, p. 75) suggests that qualitative research questions are suitable for answering questions about: (a) the *meanings* attributed by participants to situations, events, behaviours and activities; (b) the influence of *context* (e.g. physical, social, temporal, interpersonal) on participants' views, actions and behaviours; and (c) the *processes* by which actions, behaviours, situations and outcomes emerge.

Whilst in quantitative research, the research questions (or hypotheses to be tested) typically drive the research and are determined at the outset, in qualitative research a more iterative process occurs (Light *et al*., 1990, p. 19). Here the researcher may have an initial set of research purposes, or even questions, but these may change over the course of the research, as the researcher finds out more about the research setting, participants, context and phenomena under investigation, i.e. deciding research questions is not a once-and-for-all affair. This is not to say that qualitative research is an unprincipled, aimless activity; rather it is to say that, whilst the researcher may have clear purposes, she is sensitive to the emergent situation in which she finds herself, and this steers the research questions. Research questions are the consequence, not the driver, of the situation and the interactions that take place within it. It is important for the qualitative researcher to ask the right questions rather than to ask about what turn out to be irrelevancies to the participants. As Tukey (1962, p. 13) remarked, it is better to have approximate, inexact or imprecise answers to the right question than to have precise answers to the wrong question. The qualitative researcher has to be sensitized to the emergent key features of a situation before firming up the research questions.

Stage 3: addressing ethical issues

Deyle *et al*. (1992, p. 623) and Hammersley and Traianou (2012) identify several critical ethical issues that need to be addressed in the research: how does one present oneself in the field? As whom does one present oneself? How ethically defensible is it to pretend to be somebody that you are not in order to gain knowledge that you would otherwise not be able to acquire or to obtain and preserve access to places which otherwise you would be unable to secure or sustain.

The issues here are several. First, there is the matter of *informed consent* (to participate and for disclosure), whether and how to gain participant assent (see also LeCompte and Preissle, 1993, p. 66). Hammersley and Traianou (2012) comment that the researcher must respect the autonomy of the participants and this means gaining informed consent and, where appropriate, regarding participants as equals in the research project (they also note that researchers studying certain groups, e.g. paedophiles, rapists, elite power groups (p. 82) may not wish to regard them as equals). They note that consideration has to be given to who gives consent, and on behalf of whom, and for what, and what 'fully informed' means (see Chapter 7 of the present volume).

Gaining consent also uncovers another consideration, namely *covert* or *overt* research. On the one hand there is a powerful argument for informed consent. However, the more participants know about the research the less naturally they might behave (Hammersley and Traianou, 2012, p. 108), and naturalism is a key criterion of the naturalistic paradigm.

Mitchell (1993) catches the dilemma for researchers in deciding whether to undertake overt or covert research. The issue of informed consent, he argues, can lead to the selection of particular forms of research – those where researchers can control the phenomena under investigation – thereby excluding other kinds of research where subjects behave in less controllable, predictable, prescribed ways, indeed where subjects may come in and out of the research over time.

Mitchell argues that in the real social world access to important areas of research is prohibited if informed consent has to be sought, for example, in researching those on the margins of society or the disadvantaged. It is to the participants' own advantage that secrecy is maintained as, if it is not, important work may not be done and 'weightier secrets' (1993, p. 54) may be kept that are of legitimate public concern. Mitchell makes a powerful case for secrecy, arguing that informed consent may excuse social scientists from the risk of confronting powerful, privileged and cohesive groups

who wish to protect themselves from public scrutiny. Secrecy and informed consent are moot points.

Patrick (1973) indicates this point sharply when, as an ethnographer of a Glasgow gang, he was witness to a murder; the dilemma was clear – to report the matter (and thereby to act legally but 'blow his cover', consequently endangering his own life) or to stay as a covert researcher, thereby breaking the law. As Ary *et al*. (2002) remark, researchers may obtain knowledge of unforeseen illicit activities, or even be part of those, and this raises ethical dilemmas for them (p. 437). The researcher, then, has to consider her loyalties and responsibilities (LeCompte and Preissle, 1993, p. 106), for example, what is the public's right to know and what is the individual's right to privacy? Researchers must decide 'whose side are we on' (Becker, 1967).

In addition to the issue of overt or covert research, LeCompte and Preissle (1993) indicate that the problems of *risk* to, and *vulnerability* of, subjects must be addressed; steps must be taken to prevent risk or harm to participants (non-maleficence – the principle of *primum non nocere*) (cf. Hammersley and Traianou, 2012; see Chapter 7 of this volume). Bogdan and Biklen (1992, p. 54) extend this to include issues of embarrassment as well as harm to those taking part (e.g. harm from physical or psychological pain, material damage, damage to a project in which people are involved, damage to reputation or status) (Hammersley and Traianou, 2012, p. 62). Bettez (2015) asks what to do with information from one participant that could be emotionally painful for another.

The question of vulnerability is present when participants in the research have their freedom to choose limited, for example, by dint of their age, health, social constraints, by dint of their lifestyle, social acceptability, experience of being victims (e.g. of abuse, of violent crime) (Bogdan and Biklen, 1992, p. 107). As the authors comment, participants rarely initiate research, so it is the responsibility of the researcher to protect them.

Ethical issues concern both those being researched and the researcher. As we mention in Chapter 13, research can also take its toll on the researcher, not only in terms of the sensitivity or nature of the topic but in terms of the process of undertaking the enquiry itself, which may be stressful, emotional and disturbing (Dickson-Swift *et al*., 2008, 2009; Blix and Wettergren, 2015; Emerald and Carpenter, 2015). Emerald and Carpenter (2015) note that researchers often downplay the emotional and personal risk of the research, in which they may feel vulnerable and exposed (p. 744). They must be aware of, and reflexive about, the emotional signals they may be obtaining about themselves in undertaking the research, and Blix and Wettergren (2015) note that this can particularly feature when gaining and maintaining access to the field and in building trust. Whilst the emotions of the researcher may, indeed, become part of the research data, this does not obviate the ethical concern of ensuring that the research does not harm the researcher. Emotional self-management is an issue (cf. Hochschild (2012) on 'emotion work').

A standard protection for participants is often the guarantee of *confidentiality* and *privacy*, withholding participants' real names and other identifying characteristics. The authors contrast this with anonymity, where identity is withheld because it is genuinely unknown (p. 106). Issues of identifiability and traceability are raised, and participants might be able to identify themselves in the research report, though others may not be able to identify them. A related factor here is the *ownership* of the data and the results, the control of the release of data (and to whom, and when) and what rights respondents have to veto the research results.

Positionality addresses relationships; it is an ethical matter. Relationships between researcher and the researched are rarely symmetrical in terms of power; it is often the case that those with more power, information and resources research those with less (Hammersley and Traianou, 2012, p. 12). Bettez (2015) notes that research knowledge, how it is produced, understood, evaluated and used, is affected by, or refracted through, 'positionalities' – how we see ourselves and others – which are influenced by cultural values, beliefs, ascribed and achieved social position, status, gender, race, sexuality, insider and outsider status etc. She argues for 'communion' in qualitative research: meaningful connection between all participants (including the researcher) in a spirit of mutual and shared equality, inclusion, respect, humanity and dignity.

Bogdan and Biklen (1992, p. 54) add to this discussion the need to respect participants as subjects, not simply as research objects to be used and then discarded. It is important for researchers to consider the parties, bodies, practices that might be interested in, or affected by, the research and the implications of the answers to these questions for the conduct, reporting and dissemination of the inquiry (Mason, 2002, p. 41). This extends to exiting the research (Ary *et al*., 2002), as the researcher may have built up strong relationships with the participants over the course of the research, and indeed is likely to have built up friendships which cannot be severed simply because the research has finished. The researcher performs a balancing act, as such friendships may develop *during* the research, and this

raises issues of mutual trust in reporting the results. The issue also concerns reciprocity and respect: how do the participants benefit from the research?

We address ethics in Chapters 7 and 8 and we advise readers to refer to these.

Stage 4: deciding the sampling

In an ideal world the researcher would be able to study a group in its entirety: a population. This was the case in Goffman's (1968) work on 'total institutions', for example, hospitals, prisons and police forces (see also Chapter 35). It was also the practice of anthropologists who were able to explore specific isolated communities or tribes. That is rarely possible nowadays because such groups are no longer isolated or insular. Hence the researcher is faced with the issue of sampling, that is, deciding which people it will be possible to select to represent the wider group (however defined). The researcher has to decide the groups for which the research questions are appropriate, the contexts which are important for the research, the time periods needed and the possible issues and artefacts of interest to the investigator. This takes the discussion beyond conventional notions of sampling.

In several forms of research, sampling is fixed at the start of the study, though there may be attrition of the sample through 'mortality' (e.g. people leaving the study), and this is problematic. Ethnographic research regards this as natural rather than irksome. People come into and go from the study. This impacts on the decision whether to have a synchronic investigation at a single point in time, or a diachronic study where events and behaviour are monitored over time to allow for change, development and evolving situations. In ethnographic inquiry sampling is recursive and ad hoc rather than fixed at the outset; it changes and develops over time. Let us consider how this might happen.

LeCompte and Preissle (1993, pp. 82–3) point out that ethnographic methods rule out statistical sampling, for a variety of reasons:

- the characteristics of the wider population are unknown;
- there are no straightforward boundary markers (categories or strata) in the group;
- generalizability, a goal of statistical methods, is not necessarily a goal of ethnography;
- characteristics of a sample may not be evenly distributed across the sample;
- only one or two subsets of a characteristic of a total sample may be important;
- researchers may not have access to the whole population;
- some members of a subset may not be drawn from the population from which the sampling is intended to be drawn.

Hence other types of sampling are required. A criterion-based selection requires the researcher to specify in advance a set of attributes, factors, characteristics or criteria that the study must address. The task is to ensure that these appear in the sample selected (the equivalent of a stratified sample). There are other forms of sampling (see Chapter 12) that are useful in ethnographic research (Patton, 1980; Guba and Lincoln, 1989; Bogdan and Biklen, 1992, p. 70; LeCompte and Preissle, 1993, pp. 69–83; Ezzy, 2002), such as:

- convenience sampling (opportunistic sampling, selecting from whomever happens to be available);
- critical-case sampling (e.g. people who display the issue or set of characteristics in their entirety or in a way that is highly significant for their behaviour). This is done in order to permit maximum applicability to others: if the information holds true for critical cases (e.g. cases where all of the factors sought are present), then it is likely to hold true for others;
- extreme-/deviant-case sampling (the norm of a characteristic is identified, then the extremes of that characteristic are located and, finally, the bearers of that extreme characteristic are selected). This is done in order to gain information about unusual cases that may be particularly troublesome or enlightening;
- typical-case sampling (where a profile of attributes or characteristics that are possessed by an 'average', typical person or case is identified, and the sample is selected from these conventional people or cases). This is done in order to avoid rejecting information on the grounds that it has been gained from special or deviant cases;
- unique-case sampling, where cases that are rare, unique or unusual on one or more criteria are identified, and sampling takes places within these. Here whatever other characteristics or attributes a person might share with others, a particular attribute or characteristic sets that person apart;
- reputational-case sampling, a variant of extreme-case and unique-case sampling, where a researcher chooses a sample on the recommendation of experts in the field;
- snowball sampling: using the first interviewee to suggest or recommend other interviewees, and so on;
- maximum variation sampling. This is done in order to document the range of unique changes that have emerged, often in response to the different conditions to which participants have had to adapt.

It is useful if the aim of the research is to investigate the variations, range and patterns in a particular phenomenon or phenomena;
- intensity sampling: according to the intensity with which the features of interest are displayed or occur;
- sampling politically important or sensitive cases. This can be done to draw attention to the case;
- convenience sampling. This saves time and money and spares the researcher the effort of finding less amenable participants.

One can add to this list, from Miles and Huberman (1994, p. 28):

- homogeneous sampling (which focuses on groups with similar characteristics);
- theoretical sampling (in grounded theory, discussed below, where participants are selected for their ability to contribute to the developing/emergent theory);
- confirming and disconfirming cases (akin to extreme- and deviant-case sampling), in order to look for exceptions to the rule, which may lead to the modification of the rule;
- random purposeful sampling (when the potential sample is too large, a smaller sub-sample can be used which still maintains some generalizability);
- stratified purposeful sampling (to identify subgroups and strata);
- criterion sampling (all those who meet some stated criteria for membership of the group or class under study);
- opportunistic sampling (to take advantage of unanticipated events, leads, ideas, issues).

Miles and Huberman make the point that these strategies can be used in combination as well as in isolation, and that using them in combination contributes to triangulation.

Patton (1980, p. 181) and Miles and Huberman (1994, pp. 27–9) also note the dangers of convenience sampling, arguing that, being 'neither purposeful nor strategic' (Patton, 1980, p. 88), it cannot demonstrate representativeness even to the wider group being studied, let alone to a wider population.

Maxwell (2005, pp. 89–90) indicates four possible purposes of 'purposeful selection':

- to achieve representativeness of the activities, behaviours, events, settings and individuals involved;
- to catch the breadth and heterogeneity of the population under investigation (i.e. the range of the possible variation: the 'maximum variation' sampling discussed above);
- to examine critical cases or extreme cases that provide a 'crucial test' of theories or that can illuminate a situation in ways which representative cases may not be able to do;
- to identify reasons for similarities and differences between individuals or settings (comparative research).

He notes that methods of data collection and sampling are not a logical corollary of, nor an analytically necessary consequence of, the research questions (p. 91). Research questions and data collection are two conceptually separate activities, though, as we have mentioned earlier in this book, the researcher needs to ensure that they are mutually informing, in order to demonstrate cohesion and fitness for purpose. Methods and sampling cannot simply be cranked out, mechanistically, from research questions; rather the methods of data collection and the research questions are strongly influenced by the setting, the participants, the relationships and the research design as they unfold over time.

Lincoln and Guba (1985, pp. 201–2) suggest an important difference between 'conventional' and naturalistic research designs. In the former the intention is to focus on similarities and to be able to make generalizations, whereas in the latter the objective is informational, to provide such a wealth of detail that the uniqueness and individuality of each case can be represented. To the charge that naturalistic inquiry, thereby, cannot yield generalizations because of sampling flaws, the writers argue that this may be necessarily though trivially true, i.e. unimportant.

Patton (1980, p. 184) suggests that 'there are no rules for sample size in qualitative inquiry', with the size of the sample depending on what one wishes to know, the purposes of the research, what will be useful and credible and what can be done within the resources available, for example, time, money, people, support – important considerations for the novice researcher.

In much qualitative research, it may not be possible, or, indeed, desirable, to know in advance whom to sample or whom to include. One of the features of qualitative research is its emergent nature. Hence the researcher may only know which people to approach or include as the research progresses and unfolds (Flick, 2009, p. 125). In this case the nature of sampling is determined by the emergent issues in the study; this is 'theoretical sampling' (Glaser and Strauss, 1967, p. 45): once data have been collected, the researcher decides where to go next, in light of the analysis of the data, in order to gather more data in order to develop his or her theory (Flick, 2009, p. 118).

Ezzy (2002, p. 74) underlines the importance of 'theoretical sampling' in his comment that, unlike other forms of research, qualitative inquiries may not always commence with the full knowledge of whom to sample, but the sample is determined on an ongoing, emergent basis. Theoretical sampling starts with data and then, having reviewed these, the researcher decides where to go next to collect data for the emerging theory (Glaser and Strauss, 1967, p. 45).

In theoretical sampling, individuals and groups are selected for their potential – or hoped-for – ability to offer new insights into the emerging theory, i.e. they are chosen on the basis of their significant contribution to theory generation and development. As the theory develops, so the researcher decides whom to approach to request their participation. Theoretical sampling does not claim to know the population characteristics or to represent known populations in advance, and sample size is not defined in advance; sampling is only concluded when theoretical saturation (discussed below) is reached. We discuss this more fully in Chapter 37.

In the educational field one could imagine theoretical sampling in the following example: interviewing teachers about their morale might give rise to a theory that teacher morale is negatively affected by disruptive student behaviour in schools. This might suggest the need to sample teachers working with many disruptive students in difficult schools, as 'critical-case sampling'. However, the study finds that some of the teachers working in these circumstances have high morale, not least because they have come to expect disruptive behaviour from students with so many problems and so are not surprised or threatened by it, and because the staff in these schools provide tremendous support for each other in difficult circumstances – they all know what it is like to have to work with challenging students.

So the study decides to focus on teachers working in schools with far fewer disruptive students. The researcher discovers that it is these teachers who experience far lower morale, and she hypothesizes that this is because this latter group of teachers has higher expectations of student behaviour, such that having only one or two students who do not conform to these expectations deflates staff morale significantly, and because disruptive behaviour is regarded in these schools as teacher weakness, and there is little or no mutual support. Her theory, then, is refined, to suggest that teacher morale is affected more by teacher expectations than by disruptive behaviour, so she adopts a 'maximum variation sampling' of teachers in a range of schools, to investigate how expectations and morale are related to disruptive behaviour. In this case the sampling emerges as the research proceeds and the theory emerges; this is theoretical sampling, the 'royal way for qualitative studies' (Flick, 2004b, p. 151). Schatzman and Strauss (1973, pp. 38ff.) suggest that theoretical sampling may change sampling according to time, place, individuals and events.

The above procedure accords with Glaser's and Strauss's (1967) view that sampling involves continuously gathering data until practical factors (boundaries) put an end to data collection, or until no amendments have to be made to the theory in light of further data – their stage of 'theoretical saturation' – where the theory fits the data even when new data are gathered. Theoretical saturation is described by Glaser and Strauss (1967, p. 61) as being reached when, even when further data are used, the properties of the category in question are not developed any further. That said, the researcher has to be cautious to avoid premature cessation of data collection; it would be too easy to close off research with limited data, when, in fact, further sampling and data collection might lead to a reformulation of the theory.

An extension of theoretical sampling is 'analytic induction', a process advanced by Znaniecki (1934). Here the researcher starts with a theory (that may have emerged from the data as in grounded theory) and then deliberately proceeds to look for deviant or discrepant cases, to provide a robust defence of the theory. This accords with Popper's notion of a rigorous scientific theory having to stand up to falsifiability tests. In analytic induction, the researcher deliberately seeks data which potentially could falsify the theory, thereby giving strength to the final theory.

We are suggesting here that, in qualitative research, sampling cannot always be decided in advance on a 'once-and-for-all' basis. It may change through the stages of data collection, analysis and reporting. Data collection, analysis, interpretation and reporting and sampling do not necessarily proceed in a linear fashion; the process is recursive and iterative. Sampling is not decided a priori – in advance – but may be decided, amended, added to, increased and extended as the research progresses. Indeed, whilst sampling often refers to *people*, in qualitative research it also refers to *events*, *places*, *times*, *behaviours*, *activities*, *settings* and *processes* (cf. Miles and Huberman, 1984, p. 36).

Many researchers will conduct short-term, small-scale qualitative research (e.g. qualitative interviews) rather than extended or large-scale ethnographic research. A fundamental question for the researcher is to decide how long to stay in a situation. Too short, and

she may miss an important outcome; too long, and key features may become a blur.

For example, let us imagine a situation of two teachers in the same school. Teacher A introduces collaborative group work to a class, in order to improve their motivation for, say, learning a foreign language. She gives them a pre-test on motivation, and finds that it is low; she conducts the intervention and then, at the end of two months, gives them another test of motivation, and finds no change. She concludes that the intervention has failed. However, months later, after the intervention has finished, the students tell her that, in fact, their overall motivation to learn that foreign language had improved, but it took time for them to realize it after the intervention. Teacher B tries the same intervention, but decides to administer the post-test one year after the intervention has ended; she finds no change to motivation levels of the students, but had she conducted the post-test sooner, she would have found a difference. Timing and sampling of timing are important.

Stage 5: finding a role and managing entry into the context

This involves matters of access and permission, establishing a reason for being there, developing a role and a persona, identifying the 'gatekeepers' who facilitate entry and access to the group being investigated (see LeCompte and Preissle, 1993, pp. 100, 111). This is complex, as the researcher will be both a member of the group and yet studying that group, so it is a delicate matter to negotiate a role that will enable the investigator to be both participant and observer. The most important elements in securing access are the willingness of researchers to be flexible and their sensitivity to nuances of behaviour and response in the participants (p. 112).

De Laine (2000, p. 41) remarks that an ability to get on with people in the situation in question, and a willingness to join in with, and share experiences in, the activities in question, are important criteria for gaining and maintaining access and entry into the field. Barley and Bath (2014) note that this is a particular challenge when conducting research with young children, and they suggest that a period of 'familiarisation' is important before the research 'officially' commences, particularly as so much advice is given to children about 'stranger-danger' (p. 184). Such familiarization can help the researcher to understand the norms, rules and rituals of the field location, developing early mutual relationships of trust, establishing positionality (discussed earlier), unobtrusively collecting data, 'mapping the setting' (p. 185) and preparing for informed consent or assent.

Wolff (2004, pp. 195–6) suggests that there are two fundamental questions to be addressed in considering access and entry into the field:

1. How can the researcher succeed in making contact and securing cooperation from informants?
2. How can the researcher position herself/himself in the field so as to secure the necessary time, space, social relations to be able to carry out the research?

Flick (1998, p. 57), summarizing Wolff's work, identifies several issues concerning entering institutions for conducting research:

1. Research is always an intrusion and intervention into a social system, and, so, disrupts the system to be studied, such that the system reacts, often defensively.
2. There is a 'mutual opacity' between the social system under study and the research project, which is not reduced by information exchange between the system under study and the researcher; rather this increases the complexity of the situation and, hence, 'immune reactions'.
3. Rather than striving for mutual understanding at the point of entry, it is more advisable to recognize agreement as a process.
4. Whilst it is necessary to agree storage rights for data, this may contribute to increasing the complexity of the agreement to be reached.
5. The field under study only becomes clear when one has entered it.
6. The research project usually has nothing to offer the social system; hence no great promises for benefit or services can be made by the researcher, yet there may be no real reason why the social system should reject the researcher.

As Flick (1998, p. 57) remarks, the research will disturb the system and disrupt routines without being able to offer any real benefit for the institution.

The issue of managing relations is critical for the qualitative researcher. We discuss issues of access, gatekeepers and informants in Chapter 12. The researcher is seen as coming 'without history' (Wolff, 2004, p. 198), a 'professional stranger' (Flick, 1998, p. 59), one who has to be accepted, become familiar and yet remain distant from those being studied. Indeed Flick (p. 60) suggests four roles of the researcher: stranger, visitor, insider and initiate. The first two essentially maintain the outsider role, whilst the latter two attempt to reach into the institution from an insider's perspective. These latter two become difficult to

manage if one is dealing with sensitive issues (see Chapter 13). This typology resonates with the four roles typically cited for observers:

OUTSIDER ←				→ INSIDER
Detached as observer	Observer	Observer as participant	Participant	Complete participant

Swain (2006), discussing ethnography, suggests that researchers may have to switch roles, from being completely passive observers to being completely active participants, as the situation demands, i.e. to draw on the complete continuum of observations and roles. Participant observation is not without its debates. Mills and Morton (2013, pp. 52–3) note that, whilst some researchers advocate participant observation as enabling the researcher to get inside the workings of the institution and its members, others are more hesitant about whether a researcher should be a participant, as this might threaten the objectivity of the researcher and, anyway, being a participant takes valuable time away from the research work of the researcher.

Role negotiation, balance and trust are significant and difficult. For example, if one were to research a school, what role should one adopt: a teacher, a researcher, an inspector, a friend, a manager, a provider of a particular service (e.g. extra-curricular activities), a counsellor, a social worker, a resource provider, a librarian, a cleaner, a server in the school shop or canteen, and so on? One has to try to select a role that will provide access to as wide a range of people as possible, preserve neutrality (not being seen as on anybody's side) and enable confidences to be secured.

Role conflict, role strain and role ambiguity are to be expected in qualitative research. For example, De Laine (2000, p. 29) comments on the potential conflicts between the researcher qua researcher, therapist and friend; she indicates that diverse roles are rarely possible to plan in advance, and are an inevitable part of fieldwork, giving rise to ethical and moral problems for the researcher, and, in turn, require ongoing negotiation and resolution.

Roles change over time. Walford (2001, p. 62) reports a staged process wherein the researcher's role moved through five phases: newcomer, provisional acceptance, categorical acceptance, personal acceptance and imminent migrant. He also reports (p. 71) that it is almost to be expected that managing different roles not only throws the researcher into questioning his/her ability to handle the situation, but brings considerable emotional and psychological stress, anxiety and feelings of inadequacy. This is thrown into sharp relief when researchers have to conceal information, take on different roles in order to gain access, retain neutrality, compromise personal beliefs and values, and handle situations where they are seeking information from others but not divulging information about themselves. Walford suggests that researchers may have little opportunity to negotiate roles and manoeuvre roles, as they are restricted by the expectations of those being researched.

A related issue is the timing of the point of entry, so that researchers can commence the research at an appropriate time (e.g. before the start of a programme, at the start of a programme, during a programme, at the end of a programme, after the end of a programme).

Further, the ethnographer seeks acceptance into the group, which engages matters of dress, demeanour, persona, age, colour, religion, ethnicity, empathy and identification with the group, language, accent, argot and jargon, willingness to become involved and to take on the group's values and behaviour etc. (see Patrick's (1973) study of a Glasgow gang). The researcher, then, must be sensitive to the significance of 'impression management' (Hammersley and Atkinson, 1983, pp. 78ff.). In covert research these factors take on added significance, as one slip could 'blow one's cover' (Patrick, 1973).

Lofland (1971) suggests that the field researcher should attempt to adopt the role of the 'acceptable incompetent', balancing intrusion with knowing when to remain apart. Such balancing is an ongoing process. Hammersley and Atkinson (1983, pp. 97–9) suggest that researchers also have to handle the management of 'marginality': they are in the organization but not of it. They comment that 'the ethnographer must be intellectually poised between "familiarity" and "strangeness", while socially he or she is poised between "stranger" and "friend"', and that this management of several roles, not least the management of marginality, can engender 'a continual sense of insecurity' (p. 100).

Gaining access and entry is a process that unfolds over time rather than a once-and-for-all matter (Walford, 2001, p. 31), as setbacks, delays and modifications can occur and have to be expected in gaining entry to qualitative research sites.

Stage 6: finding informants

This involves identifying those people who have the knowledge about the group, issue or institution being studied. This places the researcher in a difficult position, for she has to be able to evaluate key informants, to decide:

- whose accounts are more important than others;
- which informants are competent to pass comments;

- which are reliable;
- what the statuses of the informants are;
- how representative are the key informants (of the range of people, of issues, of situations, of views, of status, of roles, of the group);
- how to see the informants in different settings;
- how knowledgeable informants actually are – do they have intimate and expert understanding of the situation;
- how central to the organization or situation the informant is (e.g. marginal or central);
- how to meet and select informants;
- how critical the informants are as gatekeepers to other informants, opening up or restricting entry to people;
- the relationship between the informant and others in the group or situation being studied.

Selecting informants and engaging with them is challenging; LeCompte and Preissle (1993, p. 95), for example, suggest that the first informants that an ethnographer meets might be self-selected people who are marginal to the group, who have a low status and who, therefore, might be seeking to enhance their own prestige by being involved with the research. Lincoln and Guba (1985, p. 252) argue that the researcher must be careful to use informants rather than informers, the latter possibly having 'an axe to grind'. Researchers who are working with gatekeepers, they argue, will be engaged in a constant process of bargaining and negotiation.

A 'good' informant, Morse (1994, p. 228) declares, is one who has the necessary knowledge, information and experience of the issue being researched, is capable of reflecting on that knowledge and experience, has time to be involved in the project, is willing to be involved in the project and, indeed, can provide access to other informants. An informant who fulfils all of these criteria he termed a 'primary informant'. Morse also cautions that not all these features may be present in the informants, but that they may still be useful for the research, though the researcher would have to decide how much time to spend with these 'secondary' informants (those who meet some but not all of the selection criteria).

Stage 7: developing and maintaining relations in the field

This involves addressing interpersonal and practical issues, for example:

- building participants' confidence in the researcher;
- developing rapport, trust, sensitivity and discretion;
- handling people and issues with which the researcher disagrees or finds objectionable or repulsive;
- being attentive and empathizing;
- being discreet;
- deciding how long to stay. Spindler and Spindler (1992, p. 65) suggest that ethnographic validity is attained by having the researcher *in situ* long enough to see things happening repeatedly rather than just once, that is to say, observing regularities.

LeCompte and Preissle (1993, p. 89) suggest that fieldwork, particularly because it is conducted face-to-face, raises challenges and questions that are less significant in research that is conducted at a distance, for example: (a) how to communicate meaningfully with participants; (b) how they and the researcher might be affected by the emotions evoked in one another, and how to handle these; (c) differences and similarities between the researcher and the participants (e.g. personal characteristics, power, resources), and how these might affect relationships between parties and the course of the research; (d) the researcher's responsibilities to the participants (qua researcher and member of their community), even if the period of residence in the community is short; (e) how to balance responsibilities to the community with responsibilities to other interested parties.

Rapport

Critically important in this area is the maintenance of trust and rapport (De Laine, 2000, p. 41), showing interest, assuring confidentiality (where appropriate) and avoiding being judgemental. De Laine adds to these (p. 97) the ability to tolerate ambiguity, to keep self-doubt in check, to withstand insecurity and to be flexible and accommodating. Such features cannot be encapsulated in formal agreements, but they are the lifeblood of effective qualitative enquiry. They are process matters.

Qualitative research recognizes that relationships emerge over time, they are not a one-off affair or in which access is negotiated and achieved on a once-and-for-all basis; rather, relationships, trust, intimacy, reciprocity, intrusion, consideration and access have to be constantly negotiated, renegotiated and agreed as time, relationships and events move on, as in real life (De Laine, 2000, pp. 83–5). In this context Maxwell (2005, p. 83) suggests that 'rapport' is problematic in discussing relationships, as it is not a unitary concept concerning its *amount* or *degree* (indeed one may have too much or too little of it) (Seidman, 1998, pp. 80–2), but its nature and kind changes over time, as people and events evolve.

Rapport and relationships influence data collection, sampling and research design (Maxwell, 2005, p. 83). Indeed, in longitudinal qualitative research, Thomson and Holland (2003, p. 235) report that maintaining and sustaining positive relationships over time can contribute significantly to lower attrition rates of participants and researchers (and attrition is a problem in longitudinal research as people move out of the area, leave as they grow older, lose contact, become too busy and so on; p. 241). Similarly, Gordon and Lahelma (2003, p. 246), researching the transition of participants from being secondary school students into becoming adults, comment that maintaining rapport is a critical factor in longitudinal ethnographic research. Rapport, they aver (p. 248), is signified in attention to non-verbal communication as well as in the sensitive handling of verbal communication.

Rapport is not easy to maintain: for example, Bettez (2015) records the dilemma when maintaining rapport with one participant might negatively affect rapport with another or with readers, and another situation where a participant in a powerful, oppressive position may not want to be reported as such, or where a family may not wish to be portrayed in a particular way as it would affect their standing in the community, i.e. where the researchers and the participants do not agree about the reporting.

Rapport is often overlaid with power relations. For example, Swain (2006, p. 205) comments that, as an adult conducting an ethnography with junior school children, he felt obliged, at times, to take the 'adult', controlling position in the research, and that he could not act as a young child, indeed that the children would find it odd if he did (p. 207). He was not a child – he was older, taller, had a deeper voice and dressed differently, but he gave the children freedom to respond to his questions as they wished. That said, he commented that he tried to adopt a role that made it clear to the children that he was not a teacher.

The issue here is that the data-collection process is itself socially situated; it is neither a clean, antiseptic activity nor always a straightforward negotiation.

Stage 8: data collection *in situ*

The qualitative researcher can use a variety of techniques for gathering information. There is no single prescription for which data-collection instruments to use; rather the issue here is of 'fitness for purpose' because, as mentioned earlier, the ethnographer is a methodological omnivore. Some qualitative research can be highly structured, with the structure being determined in advance of the research (pre-ordinate research), for example in order to enable comparisons to be made – similarities and differences (e.g. Miles's and Huberman's (1984) cross-site analysis of several schools).

Less structured approaches to qualitative research enable specific, unique and idiographic accounts to be given, in which the research is highly sensitive to the specific situation, the specific participants, the relationships between the researcher and the participants (Maxwell, 2005, p. 82), and the emergent, most suitable ways of conducting the data analysis.

For data collection the researcher can use field notes, participant observation, journal notes, interviews, diaries, life histories, artefacts, documents, video recordings, audio recordings etc. Several of these are discussed elsewhere in this book. Lincoln and Guba (1985, p. 199) distinguish between 'obtrusive' methods (e.g. interviews, observation, non-verbal language) and 'unobtrusive' methods (e.g. documents and records), on the basis of whether another human typically is present at the point of data collection.

Field notes can be written both *in situ* and away from the situation. They contain the results of observations, analysis, researchers' comments and self-memos (cf. Mills and Morton, 2013, chapter 4). The nature of observation in ethnographic research is discussed fully in Chapter 26 of the present volume. Accompanying observation techniques are interviews, documentary analysis and life histories (discussed in Chapters 25 and 16). A popularly used interview technique employed in qualitative research is the semi-structured interview, where an interview schedule (list of items, questions, prompts and probes) is prepared that is sufficiently open-ended to enable the contents to be re-ordered, digressions and expansions made, new avenues to be included and further probing to be undertaken. Carspecken (1996, pp. 159–60) describes how such interviews can range from the interrogator giving bland encouragements, 'non-leading' leads, active listening and low-inference paraphrasing to medium- and high-inference paraphrasing. In interviews, the researcher might wish to further explore some matters arising from observations. In naturalistic research, validity in interviews include honesty, depth of response, richness of response and commitment of the interviewee (Oppenheim, 1992).

Lincoln and Guba (1985, pp. 268–70) propose several purposes for interviewing, including: *present constructions* of events, feelings, persons, organizations, activities, motivations, concerns, claims, etc.; *reconstructions* of past experiences; *projections* into the future; *verifying, amending and extending data*. Silverman (1993, pp. 92–3) adds that interviews in qualitative research are useful for: (a) gathering facts;

(b) accessing beliefs about facts; (c) identifying feelings and motives; (d) commenting on the standards of actions (what could be done about situations); (e) exploring present or previous behaviour; (f) eliciting reasons and explanations.

Lincoln and Guba (1985) emphasize that the planning of the conduct of the interview is important, including the background preparation, the opening of the interview, its pacing and timing, keeping the conversation going and eliciting knowledge, and rounding off and ending the interview. It is important for careful consideration to be given to the several stages of the interview. For example, at the planning stage, attention will need to be given to the number of interviews per interviewer, duration, timing, frequency, setting/location, number of people in a single interview situation (e.g. individual or group interviews) and respondent styles (LeCompte and Preissle, 1993, p. 177). At the implementation stage the conduct of the interview will be important, for example, responding to interviewees, prompting, probing, supporting, empathizing, clarifying, crystallizing, exemplifying, summarizing, avoiding censure, accepting. At the analysis stage there are several considerations, for example: the ease and clarity of communication of meaning; the interest levels of the participants; the clarity of the question and the response; the precision (and communication of this) of the interviewer; how the interviewer handles questionable responses (e.g. fabrications, untruths, claims made).

The qualitative interview tends to move away from a pre-structured, standardized format and towards an open-ended or semi-structured format (see Chapter 25), which enables respondents to project their own ways of defining the world. It permits flexibility rather than fixity of sequence of discussions, allowing participants to raise and pursue issues and matters that might not have been included in a pre-devised schedule (Denzin, 1970; Silverman, 1993).

The use of interviews is not automatic for qualitative research. Some participants may find it alien to their culture; they may feel uncomfortable with interviews, or indeed with any such formal verbal communication (Maxwell, 2005, p. 93). The qualitative researcher has to find a culturally appropriate and culturally sensitive way of gathering data. Maxwell (echoing Whyte, 1993, p. 303, discussed in Chapters 12 and 13) cites sensitive research (heroin users) which indicates that it is unwise or inappropriate to ask too many questions, and that conducting formal interviews is an alienating activity, better replaced by informal conversations and field notes.

In addition to interviews, Lincoln and Guba (1985) discuss data collection from non-human sources, including:

- documents and records (e.g. archival records, private records). These have the attraction of being always available, often at low cost, and being factual. On the other hand, they may be unrepresentative or selective, they may lack objectivity, may be of unknown validity and may possibly be deliberately deceptive (see Finnegan, 1996; see also Chapter 16);
- unobtrusive informational residues. These include artefacts, physical traces and a variety of other records. Whilst they frequently have face validity, and whilst they may be simple and direct, gained by non-interventional means (hence reducing the problems of reactivity), they may also be very heavily inferential, difficult to interpret and may contain elements whose relevance is questionable.

Qualitative data collection is not hidebound to a few named strategies; it is marked by eclecticism and fitness for purpose. It is not to say that 'anything goes' but that 'use what is appropriate' is sound advice. Mason (2002, pp. 33–4) advocates integrating methods, for several reasons:

- to explore different elements or parts of a phenomenon, ensuring that the researcher knows how they interrelate;
- to answer different research questions;
- to answer the same research question but in different ways and from different perspectives;
- to give greater or lesser depth and breadth to analysis;
- to triangulate – corroborate – by seeking different data about the same phenomenon.

She argues that integration can take many forms, and she suggests that researchers should consider whether the data are to complement each other, to be combined, grouped and aggregated, and to contribute to an overall picture. She also argues that it is important for the data to complement each other *ontologically*, to be ontologically consistent (p. 35). Added to this, integration must be in an *epistemological* sense, i.e. where the data emanate from the same, or at least complementary, epistemologies, and whether they are based on 'similar, complementary or comparable assumptions' (p. 36) about what researchers can legitimately constitute as evidential knowledge. Finally, she argues that integration must occur at the level of *explanation*. By this she means that the data from different sources and methods

must be able to be combined into a coherent, convincing and relevant explanation and argument (p. 36).

Data collection also relates to sampling. For example, in qualitative or ethnographic interviews, though the researcher may wish to include a range of participants, in fact some of those participants may be shy, inarticulate, marginalized, dominated, introverted, overwhelmed or fearful in the presence of others or of being censured, or uninterested in participating (Swain, 2006, p. 202). In these circumstances the researcher may have to use alternative methods of gathering data, such as observation. Miller and Dingwall (1997) point out that an interview may be very unsettling for some participants, being too formal or unnatural; it is not the same as a conversation, and some participants may not 'open up' in a non-conversational situation. We discuss interviews and interviewing in Chapter 25.

Stage 9: data collection outside the field

In order to make comparisons and to suggest explanations for phenomena, researchers might find it useful to go beyond the confines of the groups in which they occur. That this is a thorny issue is indicated in the following example. Two students are arguing violently and physically in a school. At one level it is simply a fight between two people. However, this is a common occurrence between these two students as they are neighbours outside school and they don't enjoy positive, amicable relations as their families are frequently feuding. The two households have been placed next door to each other by the local authority because it has taken a decision to keep together families who are very poor at paying for local housing rent (i.e. a 'sink' estate). The local authority has taken this decision because of a government policy to keep together disadvantaged groups so that targeted action and interventions can be more effective, thus meeting the needs of whole communities as well as individuals.

The issue here is: how far out of (or indeed inside) a micro-situation does the researcher need to go to understand that micro-situation (Morrison, 2009, p. 7), for example, the individual, familial, neighbourhood, local government or national government level?

Stage 10: data analysis

Though we devote six chapters specifically to qualitative data analysis later in this book (Part 5), there are some preliminary remarks that we make here. Data analysis involves organizing, accounting for and explaining the data; in short, making sense of data in terms of participants' definitions of the situation, noting patterns, themes, categories and regularities. Typically in qualitative research, data analysis commences during the data-collection process. There are several reasons for this, discussed below.

At a practical level, qualitative research rapidly amasses huge amounts of data, and early analysis reduces the problem of data overload by selecting significant features for future focus. Miles and Huberman (1984) suggest that careful data display is an important element of data reduction and selection. 'Progressive focussing', according to Parlett and Hamilton (1976), starts with the researcher taking a wide-angle lens to gather data, and then, by sifting, sorting, reviewing and reflecting on them, the salient features of the situation emerge. These are then used as the agenda for subsequent focusing. The process is like funnelling from the wide to the narrow.

Maxwell (2005, p. 95) argues for data analysis not only to be built into the design of qualitative research, but to start as soon as each stage or round of data collection happens, or as soon as any data have been collected, i.e. without waiting for the next stage, round or piece of data to have taken place. He cites the analogy of the fox having to keep close to the hare: keeping the collection and the analysis close together ensures that the researchers can keep close to changes and their effects. He suggests that data analysis commences with careful reading and re-reading of the data, then constructing memos, categorizations (e.g. coding into organizational, substantive – descriptive – and theoretical categories (e.g. related to prior theory, 'etic' categories, grounded theory)) and thematic analysis, and 'connecting strategies' such as narrative analysis (p. 96) and vignettes, discourse analysis and profiles (p. 98) that set the data in context and indicate relationships between different parts of the data such that the integrity – the wholeness – of the original context is preserved (p. 98), rather than the fracturing and regrouping of the data that can occur in a coding exercise.

Analytical memos, including striking observations and comments, enable researchers to make connections between observations, analysis and literature (Mills and Morton, 2013, p. 122). They act as a record, a reminder, a focus, a conjecture, a tentative explanation and a suggestion for future steps to take in the research.

At a theoretical level, a major feature of qualitative research is that analysis commences early on in the data-collection process so that theory generation can happen (LeCompte and Preissle, 1993, p. 238). LeCompte and Preissle (1993, pp. 237–53) advise researchers to: (a) set out the main outlines of the phenomena that are under investigation; then (b) assemble chunks or groups of data, putting them together to make a coherent whole (e.g. through writing summaries of what has been found); then (c) painstakingly take

apart their field notes, matching, contrasting, aggregating, comparing and ordering notes made. The intention is to move from description to explanation and theory generation.

Thomson and Holland (2003, p. 236) suggest that, in longitudinal qualitative research, data analysis should be both cross-sectional (in order to discover the discourses and themes at work in the construction of identities and interpretations at a particular point in time) and longitudinal (in order to chart the development of narrative(s) over time). However, they also recognize that cross-sectional approaches and longitudinal approaches may sit together uncomfortably, as the former chops up and reassembles text from different participants in order to present themes at one moment in time, whilst the latter seeks individual narratives that require the continuity that only emerges over time and within individuals (p. 239).

Longitudinal research that uses ethnographic techniques (e.g. life histories) can also be used to chart transitions in participants, for example, from primary to secondary school, from secondary school to university, from school to work, from childhood to adulthood etc. Gordon and Lahelma (2003) comment that in such research, the reflexivity of the participants can increase over time, and that sensitivity and rapport (discussed earlier) are key elements for success. Indeed the authors go further, to argue that as the research develops over time, so does the obligation to demonstrate reciprocity in the relationships between researcher(s) and participants, so that, just as the participants give information, so the researcher has an ethical obligation to ensure that the research offers something positive, in return, to the participants. This need not necessarily mean a material incentive or reward; it could mean an opportunity for the participants to reflect on their own situation, to learn more about themselves and to support their development (p. 249). In this case reflexivity is not confined to the researcher, but extends to the participants as well (p. 252).

We discuss cross-sectional and longitudinal studies (surveys) in Chapter 17.

Thomson and Holland (2003) indicate the frustration and intimidation that early analysis in longitudinal research can cause for researchers, as there is never complete closure on data analysis, as 'the next round of data' can challenge earlier interpretations made by researchers. Indeed they question when the right time is to commence writing up or make interpretations.

In addition to the challenge of continual openness to interpretation as qualitative research unfolds is the related issue of whose views/voices one includes in the data analysis, given that, in the interests of practicality, it may not be possible to include everyone's voice, even though the canons of validity in qualitative research might call for multiple voices to be heard. Eisenhart (2001, p. 19) points out that researchers all too easily can privilege some voices at the expense of others and that the express, beneficent intention of protecting some participants can have the effect of silencing them. How will the researcher present different, even conflicting voices, accounts or interpretations? What are the politics surrounding inclusion and exclusion of voices? We return to this issue in Part 5 on qualitative data analysis.

For clarity, the process of data analysis can be portrayed in a sequence of seven steps which are set out here and addressed in subsequent pages (Figure 15.4).

Step 1: establish units of analysis of the data, indicating how these units are similar to and different from each other

The criterion here is that each unit of analysis (category – conceptual, actual, classification element, cluster, issue) should be as discrete as possible whilst retaining fidelity to the integrity of the whole, i.e. that each unit must be a fair rather than a distorted representation of the context and other data. The creation of units of analysis can be done by ascribing *codes* to the data (Miles and Huberman, 1984). This is akin to the process of 'unitizing' (Lincoln and Guba, 1985, p. 203).

Step 2: create a 'domain analysis'

A domain analysis involves grouping together items and units into related clusters, themes and patterns, a domain being a category which contains several other categories.

Step 3: establish relationships and linkages between the domains

This process ensures that the data, their richness and 'context-groundedness' are retained. Linkages can be found by identifying confirming cases, by seeking 'underlying associations' (LeCompte and Preissle, 1993, p. 246) and connections between data subsets. This helps to establish core themes, i.e. those themes which seem to underpin or to have reference made to them most frequently or most significantly in the data, or which explain a lot (Gonzales *et al*., 2008, pp. 5–6).

Step 4: make speculative inferences

This stage moves the research from description to inference. It requires the researcher, on the basis of the evidence, to posit some explanations for the situation, some key elements and possibly even their causes. It is

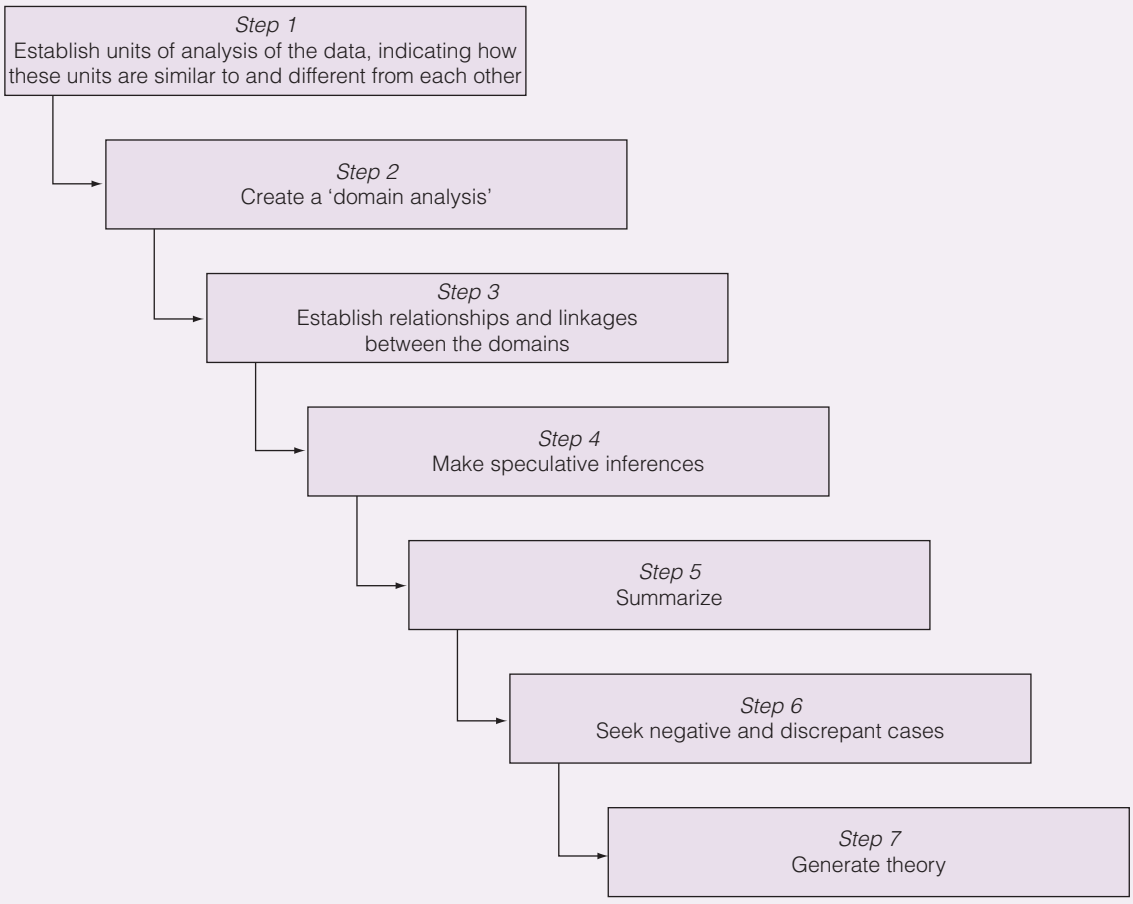


FIGURE 15.4 Seven steps in qualitative data analysis

the process of hypothesis generation or the setting of working hypotheses that feeds into theory generation.

Step 5: summarize

Here the researcher writes a preliminary summary of the main features, key issues and key concepts, constructs and ideas encountered so far in the research. We address summarizing in more detail in Chapter 33.

Step 6: seek negative and discrepant cases

In theory generation it is important to seek not only confirming cases but to weigh the significance of disconfirming cases. LeCompte and Preissle (1993, p. 270) suggest that because interpretations of the data are grounded in the data themselves, results that fail to support an original hypothesis are neither discarded nor discredited; rather, it is the hypotheses themselves that must be modified to accommodate these data. LeCompte and Preissle (1993, pp. 250–1) define a negative case as an exemplar which disconfirms or refutes the working hypothesis, rule or explanation so far. The theory that is being developed becomes more robust if it addresses and can embrace or explain negative cases, for it sets the boundaries to the theory, modifies the theory and sets parameters to the applicability of the theory.

Discrepant cases are not so much exceptions to the rule (as in negative cases) as variants of the rule (LeCompte and Preissle, 1993, p. 251). The discrepant case leads to the modification or elaboration of the construct, rule or emerging hypothesis. Discrepant case analysis requires the researcher to seek out cases for which the rule, construct or explanation cannot account or with which they will not fit, i.e. they are neither exceptions nor contradictions, they are simply different!

Step 7: generate theory

Here the theory derives from the data; it is grounded in the data and emerges from it (see Chapter 37). As Lincoln and Guba (1985, p. 205) argue, grounded theory must fit the situation that is being researched. Grounded theory is an iterative process, moving backwards and forwards between data and theory until the theory fits the data. This breaks the linearity of much conventional research (Flick, 1998, pp. 41, 43) in which hypotheses are formulated, sampling is decided, data are collected and then analysed and hypotheses are supported or not supported. In grounded theory, a circular and recursive process is adopted, wherein modifications are made to the theory in light of data, more data are sought to investigate emergent issues (theoretical sampling), and hypotheses and theories emerge from the data.

Lincoln and Guba (1985, pp. 354–5) urge the researcher to be mindful of several issues in analysing and interpreting the data, including: (a) data overload; (b) the problem of acting on first impressions only; (c) the availability of people and information (e.g. how representative these are and how to know if missing people and data might be important); (d) the dangers of seeking only confirming rather than disconfirming instances; (e) the reliability and consistency of the data and confidence that can be placed in the results.

Maxwell (2005, p. 108) draws attention to some important issues of validity for the qualitative data analyst, including researcher bias and reactivity. The former concerns the projection of the researcher's own values and judgements onto the situation, whilst the latter concerns the effect of the research(er) on the participants, giving rise to unreliable behaviours or changes to the natural setting (a particular problem, for example, in interviewing or observing children). Maxwell sets out a useful checklist of ways in which attention can be given to validity in qualitative research:

- *intensive, long-term involvement*, enabling the researcher to probe beneath immediate behaviours, for reducing reactivity and for revealing causal processes;
- *'rich' data*, sufficient to provide a sufficiently, revealing, varied and full picture of the phenomenon, participants and settings;
- *respondent validation*, to solicit feedback from participants on the interpretations made of, and conclusions from, the data;
- *intervention*, where the researcher intervenes formally or informally, in a small or a large way, in the natural setting in order to contribute positively to a situation (whether this is legitimate is a moot point, as it disturbs the natural setting, even though its intention might be in the interests of serving the ethical issue of 'beneficence'; see Chapter 7);
- *searching for discrepant evidence and negative cases*, in order to constitute a strong test of the theory or conclusions drawn;
- *triangulation*, in order to give reliability to the findings and data (see Chapter 14);
- *quasi-statistics*, where quasi-quantitative statements are interrogated, for example, claims that a finding is rare, extreme, unusual, typical, frequent, dominant, prevalent and so on;
- *comparison*, between groups, sub-groups, sites and settings, events and activities, times, contexts, behaviours and actions etc., to look for consistency or inconsistency, similarity or difference across these.

Swain (2006, p. 202) comments that, in writing up an ethnography or qualitative research, the researcher must exercise discipline, in that a faithful account has to be written, yet, for manageability, the level of detail on the context, emerging situation and events has to be reduced. Indeed he argues that less than 1 per cent of the collected data may feature in the final report, and that, even if all the data that were collected were included, these would constitute less than 1 per cent of everything that took place or that was experienced by the researcher. Fidelity to the detail may stand in a relation of tension to the final, necessarily selective, use of data, and care has to be given to issues of reliability and validity in such a situation.

These are significant issues in addressing reliability, trustworthiness and validity in the research (see Chapter 14). Further, the essence of this approach, that theory emerges from and is grounded in data, is not without its critics. For example, Silverman (1993, p. 47) suggests that it fails to acknowledge the implicit theories which guide research in its early stages (i.e. data are not theory-neutral but theory saturated) and that the theory might be strong on providing categorizations without necessarily explanatory potential. These caveats should feed into the process of reflexivity in qualitative research.

Maxwell (2005, pp. 115–16) also indicates that the process of data analysis, and the conclusions drawn from the data, should address generalizability, i.e. to whom the results are generalizable. Internal generalizability will indicate that the results and conclusions are generalizable to the group in question, whilst external generalizability will indicate that the results and conclusions are generalizable to the wider population beyond the group under study. He suggests that, whilst

the former may be applicable to qualitative research, the latter often may not. However, he also indicates that this by no means rules out the external generalizability of qualitative studies, as respondents themselves might have commented on the generalizability of their situation, or the researcher or readers might see similarities to other, comparable situations, constraints or dynamics, or the research might be corroborated by, or corroborate, other studies. He indicates, however, that external generalizability is not a strong feature, indeed a concern, of qualitative research.

Stage 11: leaving the field

The issue here is how to conclude the research, how to terminate the roles adopted, how (and whether) to bring to an end the relationships that have built up over the course of the research, and how to disengage from the field in ways that bring as little disruption to the group or situation as possible (LeCompte and Preissle, 1993, p. 101). De Laine (2000, p. 142) remarks that some participants may want to maintain contact after the research is over, and not to do this might create, for them, a sense of disappointment, exploitation or even betrayal.

The researcher has to consider the after-effects of leaving and take care to ensure that nobody comes to harm or is worse off from the research, even if it is impossible to ensure that they have benefited from it.

Stage 12: writing the report

Often the main vehicle for writing naturalistic research is the case study (see Chapter 19), whose 'trustworthiness' (Lincoln and Guba, 1985, p. 189) is defined in terms of credibility, transferability, dependability and confirmability (see Chapter 14). Case studies are useful in that they can provide the thick descriptions that typify ethnographic research, and can catch and portray to the reader what it is like to be involved in the situation (p. 214). As Lincoln and Guba comment (p. 359), the case study is the ideal instrument for 'emic' inquiry. They provide several guidelines for writing case studies (pp. 65–6):

- the writing should strive to be informal and to capture informality;
- as far as possible, the writing should report facts except in those sections where interpretation, evaluation and inference are made explicit;
- in drafting the report it is more advisable to opt for over-inclusion rather than under-inclusion;
- the ethical conventions of report writing must be honoured, for example, anonymity, non-traceability;
- the writer should make clear the data that give rise to the report, so the readers have a means of checking back for reliability and validity and inferences;
- a fixed completion date should be specified.

Spradley (1979) suggests a sequence of nine practical steps in writing an ethnography:

1. Select the audience.
2. Select the thesis.
3. Make a list of topics and create an outline of the ethnography.
4. Write a rough draft of each section of the ethnography.
5. Revise the outline and create subheadings.
6. Edit the draft.
7. Write an introduction and a conclusion.
8. Re-read the data and report to identify examples.
9. Write the final version.

Clearly there are several other aspects of case study reporting that need to be addressed. These are set out in Chapter 19.

The writing of a qualitative report can also consider the issue of the generalizability of the research. Whilst much qualitative research strives to embrace the uniqueness and individual idiographic features of the phenomenon and/or participants, rendering generalization irrelevant (though the study would still need to ensure that it contributes something that is worthwhile and significant for the research community), this need not preclude attention to generalization where applicable in qualitative research. Indeed one can question the value or contribution of idiographic research that does not have any generalizable function or utility (Wolcott, 1994, p. 113).

Generalization takes many forms; it is not a unitary or singular concept, and it connotes far more than the familiar terms 'transferability' (Denzin and Lincoln, 1994) or 'external validity' (Cook and Campbell, 1979). Larsson (2009, p. 27) comments that defining generalization as that which is derived by strict sampling from a defined population is often irrelevant in qualitative research. He also suggests that those single studies that seek to undermine 'universal' truths similarly do not need to aspire to be generalizable, as the single instance of falsification ('negative cases'; p. 30) may be sufficient to bring down the theory (though the case would need to be made that the 'truths' claimed to be universal in the first place as social actions may not be susceptible to universal laws of behaviour). However, he suggests three kinds of reasoning on which generalization in qualitative research might be useful:

1 Enhancing the potential for generalization by maximizing the range of a sample's characteristics in exploring a particular issue (e.g. in theoretical sampling) or phenomenon, i.e. to ensure that as many different cases or categories of an issue as possible are included in the research. Here uncommon cases have as equal a weight as the typical cases, and the variation that exists within the study should be expected to exist in the wider population, context or situation to which one wishes to generalize (p. 31). This, in turn, may require a larger sample than may be normal in qualitative research, in order to have as broad a variation and range of characteristics as possible included, and this may not be possible in some qualitative research, for example, case studies. It also assumes that the researcher will know what the maximum variation will look like, so that he or she knows when it is reached, and this, too, may not be realistic (p. 32).
2 Generalization by ensuring the similarity of contexts between that of the qualitative research and the wider contexts to which it is wished to be applied (akin to the 'transferability' criterion of Guba and Lincoln (1994)). Here Strauss and Corbin (1990, p. 267) argue that generalizability might also be replaced by 'explanatory power' in the context of the research and the wider contexts. This view of generalizability assumes that the characteristics of the wider contexts are known, and this may not be for the researcher to judge, but, rather, for the outsider readers, audiences or users of the research to make such judgements (cf. Wolcott, 1994, p. 113). Hence, Larsson (2009, p. 32) argues, the task of the researcher is to provide sufficient details and 'thick descriptions' for the audiences to come to an informed judgement about generalizability here. A problem is raised in this kind of generalizability, in deciding when and on what – and how many – criteria the contexts of the research and the wider contexts are similar and when sufficient similarity of contexts has been reached for the research to be generalizable to those wider contexts (p. 33), as the same people or kinds of people may act differently in different – or even the same – contexts.
3 Generalization by recognizing similar patterns between the research and other contexts (Larsson, 2009, pp. 33–5) in terms of, for example: theoretical constructions; themes; concepts; behaviours; assumptions made; processes; interpretations of actions, events or descriptions. Here the issue of interpretation is raised, as interpretations of one context may be very different from the interpretations made of another – however similar – context. Whether a pattern is indeed a pattern, or whether a construction is an acceptable construction, is a matter of debate and interpretation. Researchers have to be sure that the patterns between both research and the wider context are, indeed, tenable. Interpretation is an inescapable feature of qualitative research, and it is this precise matter that renders difficult the applicability of research from one context to another, because it is not the context but the *interpretation* of the context that has to be similar to that to which it is being applied. Further, one is faced with the added problem of identifying whose interpretation should stand (not only the issue of 'emic' and 'etic' research, but also whose 'etic' and 'emic' interpretations, given that there will be multiple variants of each type).

Larsson (2009, p. 36) is arguing powerfully that responsibility for generalization from qualitative research resides with the audience rather than the researcher. However, to suggest this may be to invite the view that the researcher has no special expertise to offer here; if so, then how is the research justified? Perhaps the solution to this is to regard the research, as with other kinds of research, as raising working hypotheses rather than conclusions, i.e. as 'work in progress' rather than unassailable truths.

Whilst it appears that writing comes late in the stage of the research, in fact it should be a continuous, ongoing activity, from the start to the finish of the research. Indeed Mills and Morton (2013) place the ongoing writing of an ethnography as a key, central feature of doing ethnographic work. Writing on an ongoing basis clarifies thoughts, observations, steps to take, reflections, analysis and so on. We strongly advise ethnographers to start writing from day one of their research.

15.11 Some challenges in qualitative, ethnographic and naturalistic approaches

There are several challenges in qualitative, ethnographic and natural approaches. These might affect the reliability and validity of the research, and include:

1 *The definition of the situation*: participants are asked for their definition of the situation, yet they have no monopoly on wisdom. They may be 'falsely conscious' (unaware of the 'real' situation), deliberately distorting or falsifying information, or being highly selective. Issues of reliability and validity here are addressed in Chapter 14 (see the discussions of triangulation).

2 *Reactivity* – the Hawthorne effect – the presence of the researcher, or the fact that it is 'research' can alter the situation as participants may wish to avoid, impress, direct, deny or influence the research(er). Again, this is discussed in Chapter 14. Reactivity can be addressed by careful negotiation in the field, remaining in the field for a considerable time and ensuring a careful presentation of the researcher's self.

3 The *halo effect* – where existing or given information about the situation or participants might be used in judging subsequent data or people, or may bring about a particular reading of a subsequent situation (the research equivalent of the self-fulfilling prophecy). This is an issue of reliability, and can be addressed by having a wide, triangulated database and the assistance of an external observer. The *halo effect* commonly refers to the researcher's belief in the goodness of participants (the participants have haloes around their heads!), such that the more negative aspects of their behaviour or personality are neglected or overlooked. By contrast, the *horns effect* refers to the researcher's belief in the badness of the participants (the participants have devil's horns on their heads!), such that the more positive aspects of their behaviour or personality are neglected or overlooked.

4 The *implicit conservatism* of the interpretive methodology. The kind of research described in this chapter, with the possible exception of critical ethnography, accepts the perspective of the participants and corroborates the status quo. It is focused on the past and the present rather than on the future.

5 There is the difficulty of focusing on the *familiar*, as participants (and, maybe, researchers too) may be so close to the situation that they neglect certain, often tacit, aspects of it. The task, therefore, is to make the familiar strange. Delamont (1981) suggests that this can be done by:

   - studying unusual examples of the same issue (e.g. atypical classrooms, timetabling or organizations of schools);
   - studying examples in other cultures;
   - studying other situations that might have a bearing on the situation in hand (e.g. if studying schools it might be useful to look at other similar-but-different organizations, for instance hospitals or prisons);
   - taking a significant issue and focusing on it deliberately, for example, gendered behaviour.

6 The *open-endedness and diversity* of the situations studied. The drive towards focusing on specific contexts and situations might overemphasize differences between contexts and situations rather than their gross similarity and routine features. Researchers should be as aware of regularities as of differences.

7 The *neglect of wider social contexts and constraints*. Studying situations that emphasize how highly context-bound they are might neglect broader currents and contexts – micro-level research risks putting boundaries that exclude important macro-level factors. Wider macro-contexts cannot be ruled out of individual situations.

8 The issue of *generalizability*. If situations are unique and non-generalizable, as many naturalistic principles would suggest, how is the issue of generalizability to be addressed? To which contexts will the findings apply, and what is the role and nature of replication studies (and are they necessary)?

9 How to write up *multiple realities* and explanations? How will a representative view be reached? What if the researcher sees things that are not seen by the participants?

10 Who *owns* the data and the report, and who has control over the release of the data?

Naturalistic and ethnographic research raises important, if challenging, questions for research in education.

To interview or not to interview?

Should the qualitative researcher, seeking to research a natural setting in as undisturbed a way as possible, intervene by interviewing, as interviewing is a non-natural activity, a disturbance of the natural setting? On the one hand, open-ended interviewing can find out participants' views on a situation, event, experience or phenomenon: it provides 'witness information' (Hammersley, 2013, p. 68) and involves participants in the situation. On the other hand, an interview is a contrived activity that is not part of the normal run of events for the participants but, rather, is a non-normal activity initiated by the researcher and his/her agenda, i.e. framing and shaping the situation through the researcher's eyes and asking for second-hand information in the sense of asking participants to comment on others' views in addition to their own. As we mention in Chapter 25, interviews are speech acts in their own right, not simply vehicles for collecting proxy data (cf. Atkinson and Delamont, 2006, p. 752).

Further, participants and interviews may not be genuine. Participants may withhold information (deliberately or not), distort the truth, promote their own agenda (e.g. 'position' themselves) and overlook the non-verbal interactions involved in interviews and their

transcription, to the extent that a 'radical critique' of interviews would reject them out of hand (Hammersley, 2013, pp. 69–72), for example, for being unreliable and invalid.

Hammersley (2006) also notes that the 'radical critique' of interviews raises questions of how far what is said in an interview really represents what is happening outside the interview (p. 9). Interviews, he contends, are their own context, and they shape what is said or not said (p. 9). Further, Atkinson and Delamont (2006) argue that short-stay, 'quick-fix' activities like interviews risk betraying the complexity of the social situation that qualitative research seeks to portray and understand, which can only be achieved by sustained research in the field.

Despite these challenges, this chapter has argued that qualitative research in its many forms is a very valuable approach to educational enquiry.

Useful websites for those commencing qualitative research are:

http://nsuworks.nova.edu/tqr (which gives the websites of several hundred other sites providing materials on qualitative research);
www.ukdataservice.ac.uk (the UK's Data Service, which includes qualitative data);
www.data-archive.ac.uk (the UK Data Archive);
http://gsociology.icaap.org/methods/qual.htm (a source for accessing other websites for online materials and support).

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Historical and documentary research

CHAPTER 16

Jane Martin

16.1 Introduction

The focus of this chapter is on research using historical and documentary evidence. Its aim is to convey the value of historical study, to consider the resources that may be available for the researcher and how relevant sources should be handled. The chapter goes beyond command of institutionalized, archival sources to consider the sheer diversity of sources and the ways these are affected by ideas of progress and loss. Debate and discussion will include: recent empirical trends and methodological arguments concerning the diverse material from which history is made; the availability and representativeness of 'historical data'; and the relationship between theory and 'facts'. In turn, the chapter points out some of the problems involved in doing historical and documentary research in educational settings. We finish by looking at a detailed case study of research that has made use of archival materials in an account that takes the figure of a woman educator activist as its central focus. The discussion will focus primarily on methodological approaches and strategies and their influence on the final outcome of the study.

The aspiration to discover what happened in the past and what it was like to live in the past is traditional to the nature of history. The challenge of accessing the voice of the past through time and space is something that excites historians. 'The past is a foreign country', sighs old Leo as he looks backs on his childhood with nostalgia in L. P. Hartley's influential novel *The Go-Between*, adapted as a film in 2015: 'they do things differently there'. Historians make history through the production of knowledge, explanations and interpretations of what has gone before. In discussing questions of sources and interpretations I shall endeavour to pass on the *pleasures* of working as an historian. It will be argued that the uses and limitations of historical sources can only be fully appreciated when they are understood in their social context as historical products. There is no straightforward sense in which history simply 'speaks for itself'. A *very* brief history of the study of history since the mid-nineteenth century will help show why.

16.2 Some preliminary considerations: theory and method

American historian Bonnie G. Smith looked at what happened when the practice of scientific history took root in nineteenth-century universities in Western Europe. In *The Gender of History* (1998) she argued that Enlightenment thinking influenced the making of a *professional* discipline in which empirically minded men defined themselves and their intellectual products in opposition to an older, more popular *amateur* history read for moral instruction and entertainment, which they deemed trivial. In such a context, history was no longer regarded as a branch of literature and a tradition of women's scholarship (as authors of textbooks, biographies or memoirs, translators and editors of original documents) disappeared from view. Consequently, when we imagine a great historian we automatically think of a man, we accept as natural titles like Malcolm Bradbury's *The History Man*, published in 1975, because professionalization and historical science developed at a time of separate spheres, when it was assumed a woman's place was in the home. Thus the making of a professional discourse involved Othering the scholarship, style and preferences of those without the ideological means to achieve disciplinary ascendancy.

Foundational claims over scientific, *empirical* history were intrinsic to the formation of this new, university-led discipline constructed upon a theory of knowledge that had its origins in the belief that knowledge derived from observation of the material world, the core tenets being: rigorous examination and knowledge of historical evidence, verified by references; impartial research, devoid of a priori beliefs and prejudices; an inductive method of reasoning, arising out of the sources themselves and moving from the particular to the universal (see also Chapter 1 of the present volume). By 1900 the assumptions about historical practice implicit within the study of history encouraged an approach consisting of the collecting and reading of the papers which official authorities drew up for the purposes of the conduct of their affairs, or which they

used in conducting them. To be steeped in primary or original sources, which were generated at the time of the event under consideration, helped lend intellectual authority among the historical profession. Scholarly reputations were built on the principles of the search for objective truth: the production of irrefutable, factual information located at the heart of the historical enquiry (Tosh, 2008, pp. 2–5).

Within this approach 'the facts' speak for themselves *independently* of a particular point of view. In *New Perspectives on Historical Writing*, British cultural historian Peter Burke (2001, pp. 3–6) stressed that the range of documents considered by traditional *empiricist* historians tended to be remarkably narrow. In the 1960s and 1970s many aspects of this account were energetically challenged (e.g. see Thompson, 1963; Hobsbawm, 1964; Rudé, 1964; Jones, 1971; Samuel, 1975). Social history, sometimes described as the 'history of the people' or 'history from below', emerged as an alternative to conventional political history, both in terms of its objects of interest and its belief in deep-rooted economic and social factors as agents of historical change. 'History from below' as practised by social historians used the language of class, including a Marxian approach to society that could be applied to a wide range of historical cases, as indeed Marx applied them. Revisionist studies in the history of education tried to question the Whiggish assumption that the course of historical development was an unbroken chain of ascent from a benighted past to an enlightened present. In their introduction to *Children, School and Society*, published in 1981, Anne Digby and Peter Searby expressed their dissatisfaction with a view of education from Westminster or Whitehall rather than the home or the schoolroom and their aspiration to provide a corrective to the dreary sequence of institutional growth that characterized former narrow, excessively bureaucratic histories of education.

A broadening of attention to other documentary sources was an important feature of the broad body of social history, allied with a willingness to supplement documents with other sources of evidence. For example, the revival of oral history derived from a new generation of historians steeped in the politics of the New Left, civil rights and feminism (e.g. Rowbotham, 1975; Thompson, 1978). Recovering lost voices provided a means to empower women, the working-class and minority communities, allowing them to speak for themselves, thereby contesting the national consensus, enlarging the explanatory concepts available to historians, generating new perspectives and radical critiques of education, youth and structural racism and sexism.

The 'take-off' of cultural history from the late 1980s onwards ruled out a fundamentally positivistic concern with getting at the truth, giving readers 'the facts'. While I shall not explore the minutiae of these approaches here, a *post-structural approach* to history entails remaking ourselves as readers and writers, giving us new methodological tools with which to approach the task of assessing and interpreting sources. Here meaning is looked for in a culture's language and systems of representation. Perception of empirical reality was constituted through multiple refracting perspectives: one that is constantly changing, subject to variation over time as well as in space. Thus, rather than a historical practice based on straightforward readings of state papers and official data, the same material artefacts or texts may be used, but read against the grain, looking for contested meanings and omissions. The linguistic turn opened up new ways of interpreting texts, which may contain different rhetorical strategies and voices as opposed to being written as supposedly objective, and especially the relationship between them – 'intertextuality'.

For historians, many post-structuralist topics and methods are a legacy of the work of Michel Foucault. Foucault studied what he termed 'the history of systems of thought' (e.g. Foucault, 1970). He argued that documents are not of interest because of what they tell us about the author, but because they inform us about the mechanisms through which power is exercised. That is, documents are a medium through which power is expressed. In adopting an interpretivist or discourse analysis approach, one would believe that the critical analysis of a document involves questioning why the document was produced, what is being said (overtly and covertly) and what is not being said (see also Chapter 3 of the present volume). Post-structuralism treats texts of all kinds as systems of signification (collections of signs which conform to some internal system), whose meanings can be ascertained in part by deconstruction, acknowledging that meaning can be self-referential and not entirely taken from the context in which it was produced or from authorial intent.

The collection and presentation of historical material is inevitably selective, and, to some extent at least, this selection and interpretation relies on the questions asked of the material, or the theoretical perspective which is brought to bear on any particular piece of research. No general consensual version of history is possible and monolithic accounts are unlikely to be either adequate or satisfactory. Different groups do have different interests, experiences and cultural forms, and do provide alternative definitions and accounts of these. Much of the historian's skill lies in the creative

and self-aware use of the sources from which history is made. Let us now turn to the task of assessing and interpreting those raw materials. It is important to appreciate the richness and the limitations of each type.

16.3 The requirements and process of documentary analysis

When discussing documentary sources it is conventional to differentiate between 'primary' and 'secondary'. The former encompass every kind of evidence which people have left of their past activities, produced during the period being studied. The latter discuss the period studied but are created sometime after it or in some way removed from the actual events that are the focus. Simply put, primary analysis is an interpretation of raw materials, whereas secondary analysis involves an examination of the interpretations of others. Obviously, it is possible for some sources to act as both primary and secondary sources, depending upon the exact context of the information we are interested in. We can perceive this in my study of the 'career' chances for twentieth-century women historians, which employed writings conventionally designated secondary sources, as primary sources. Thus, whereas Eileen Power's *Medieval Women* is a secondary text for a historian of the period, for me, as an example of Power's *oeuvre*, it became a primary source (Martin, J., 2014).

When we decide to use historical documents, their validity and reliability must always be held up for scrutiny. Scott (1990) identifies four potential challenges. To start we have to consider the issue of *authenticity*; who a document was written for and by (authorship); whether it constitutes a first-hand, second-hand or even more remote account; whether confidential or not, public or private, forced or voluntary and so on. At a simple descriptive level, whether the document is 'sound' or *authentic* may be challenged on several grounds – if it contains many errors, is one of many versions, is inconsistent in relation to other similar documents and in terms of 'ownership'. It is not always the case that the identity of the author is apparent.

Credibility is a second potential challenge. In other words, is the document we are analysing reliable? Is it undistorted, 'sincere' and 'accurate'? For a document to be credible, we need to be aware of the purpose of the document. Was it produced to describe events, to persuade (such as a school brochure) or to self-protect (such as a ministerial memorandum)? The purpose of a communication, then, can provide an important context for understanding its content. This is not to say that you should be suspicious of every document you encounter. However, there is always a possibility that the author of an official document did not believe what he or she recorded, while personal documents may be produced for a host of reasons, depending on the mood of the moment.

Third, we need to ask whether the document is *representative*. Are we looking at a unique view or does it represent a 'general mood of the time'? Assessing the typicality, or otherwise, of evidence centres on the two aspects of 'survival' and 'availability'. Not every document will make its way into an archive: documents have differential survival rates and those which do survive do not always provide all the information required. The answers to a great many questions are simply not available, since the necessary records either never existed or failed to survive. With respect to the UK national archive, selection of public records takes place in two stages. At the outset, records which are considered worthless are destroyed, and those which have been identified as valuable for future administrative need or future research are kept for further review when the record is fifteen to twenty-five years old. The UK's Freedom of Information Act (2000) governs access to information held by most public authorities, with two forms of exemption: 'absolute' and 'qualified'. In the case of the latter, a public interest test must be made, balancing the public interest in maintaining the exemption against the public interest in disclosing the information. A classic example is data on school exclusion, which would be subject to a 100-year rule because disclosure could cause distress to living individuals.

Finally, we need to pay attention to the *meaning* of the document. There are three aspects to this. These are: the *intended content* of a text, the *received content* of a text and the *internal meaning* of a text. Acknowledging the complications of text comprehension, Scott (1990) describes the process of understanding a text hermeneutically according to his four criteria (authenticity, credibility, typicality and meaning). To some extent a heuristic tool, the hermeneutic circle requires closing the loop whereby knowledge and understanding of the text as a whole is achieved through 'dialogue' within boundaries set by the frames of reference of the researcher and those who produced the text.

16.4 Some problems surrounding the use of documentary sources

Documents are selective in terms of the information presented. Some documents are produced with research in mind (most institutions and some individuals have deliberately sought, with an eye on the future, to generate accounts of their activities); others are produced for

personal use and are less self-conscious. Either way, the act of recording will also be informed by the social, cultural, economic and political landscape of which they are a part.

It is important to ask who created the source and how it was understood by contemporaries. When political texts are employed, intellectual historian Quentin Skinner (2002) takes this process a step further. First, by insisting that the works of political thinkers be understood within the context out of which their works were produced, and second, that they be understood as acts of rhetorical communication – to consider the intentions of the author and how they were received.

A less obvious problem is that those who work with historical documentary evidence can have both too little and too much available to them. You might never find exactly what you need; on the other hand, where, in the large amount of twentieth-century material for example, do you start? In reality, our understanding of educational history is informed by a selective reading of documents. It is highly unlikely, particularly for undergraduate students, that you will have had the time or opportunity to read *everything* about a person (for example, an education minister) or an event (for example, the imposition of a legally enforced national curriculum). Similarly, the *types* of document we read vary in terms of information and accuracy: some are very opinionated or subjective whilst others may be highly factual and descriptive.

The categorization of historical documents largely follows the dominant definitions of a particular period and if we are to read the silences we need also to look at material which gives little hint that evidence we need may be lurking inside it. There are several consequences of this approach. First of all, for much of the time one is working in the dark. Decisions about what to look at and what to ignore may be a hit-and-miss affair. Second, this kind of research is time-consuming, frustrating, often unrewarding and frequently leads to a feeling of wasted time and effort. Third, the amount of material which historical researchers often have to handle makes some kind of selective reading or sampling, whether deliberate or otherwise, inevitable. But on what basis should choices be made? The collection and presentation of raw materials is inevitably selective, and to some extent at least this selection and interpretation relies on the questions asked of the material, or the theoretical perspective which is brought to bear on any particular piece of research.

One method of selection is to focus on 'significant' events or periods, guided by information from secondary sources. One drawback of this approach is the possible risk of interpreting a set of extraordinary circumstances as being more generally applicable. It means we cannot explain change over time, or understand apparent continuities and apparent breaks in activity. Alternately, one could adopt a more random approach, for example, the selection of, say, one newspaper a month, or more systematic sampling (see Chapter 12 of the present volume) of every tenth newspaper. An obvious drawback to this method is the difficulties it presents in following up stories. The best solution to the problem may be the use of a variety of sources simultaneously so that an overall pattern begins to emerge which suggests directions for future research. There can be no 'formula' for decisions of this kind. A 'feel' for the period, the relevant questions and the sources is ultimately what guides the methodology of any historical research project, and this can only properly be gained as part of the research process itself. Ideally it may be that the research should be considered finished when all the classes of document relevant to it have been exhausted, but this is rarely a practical proposition. There is a sense, inevitably, in which the research is over whenever it is time to stop.

Quality appraisal is a never-ending process. This does *not* mean that only error-free, typical documents (from written to oral) which have a common interpretation can be used. What it *does* mean is that we must be at least aware of any challenges to reliability that may exist, or that others may interpret the information held within differently.

16.5 The voice of the past: whose account counts?

A move away from structuralism in all the social sciences is important in accounting for the increasing numbers of historians who have turned to biography in the last thirty years. In a widely read textbook on the study of history, Ludmilla Jordanova (2000, p. 41) likens biography to 'holistic history'. Barbara Finkelstein (1998), Diana Jones (1998) and Peter Figueroa (1998) all indicate clearly the considerable strength of the biographical approach for an understanding, and bringing to life, of the history of education. Educational biography offers a frame of reference within which to assess the relative power of material and ideological circumstances, the meaning of policy and practice, the utility of formal and informal schooling and the relationship between learning and teaching. Researchers utilize any form of writing that includes a construction of the self (diaries, memoirs, letters, autobiography and biography, travel writing), oral testimony, photographs and material objects (see also Plummer, 2001).

Kathleen Weiler's (1998) study of women teachers in two rural California counties from 1850 to 1950 relies heavily on oral testimony. Texts and testimonies explore the social contexts of teaching, employed to understand what teaching meant to women teachers, what it provided them and how it shaped their categories of *experience*. Building on the work of gender historian Joan Scott (1986) and influenced by Foucault, three key concepts inform Weiler's approach – *knowledge*, *language* and *subjectivity*. Used to convey the constructed quality of memory and experience, subjectivity includes the struggle and contest over identity, the process of identification and an unstable, shifting subject constructed both through dominant conceptions and resistance to those conceptions.

Peter Cunningham and Peter Gardner (2004) also used interviews as a source to help reconstruct what being a student teacher meant to various groups in early-twentieth-century England. They wanted to write a different kind of history concerned with the day-to-day experience of ordinary teachers as opposed to the administration of education. For them this was not solely a matter of restoring 'lost voices' or interlocutors; it was also about recording and interpreting events previously excluded both for contemporaries and for future generations. It was about the creation of a more accurate historical record and therefore a more 'useable' past that includes possibilities we might not even have considered because the record of the road not followed was less likely to survive (possibly destroyed by people with different priorities). Opening themselves to this recognition produced a history-writing that is wider in scope and does not just reflect the standpoint of authority.

Oral testimonies also formed the core of a project exploring work and identity in three main occupational sectors, including teachers, in twentieth-century England. Overall the team conducted forty interviews with three generations of men and women teachers: retired, mid/late working life and younger teachers or people who were new entrants, with interviews taking place in London and the south-east and the north-west of England. Many of the interviews with older people were conducted in people's homes, allowing, in these cases, for a more reflexive account of working life, whereas those spoken to in the workplace gave shorter answers to the interviewers' prompts: the pressure of a working day pre-empting any long conversations, but not precluding articulate insights into the narration of the 'teacherly self'. Whilst they all talked about 'making a difference', 'postmodern' teachers expressed themselves in terms of a kind of entrepreneurial culture, whereas earlier generations articulated the vocabulary and presuppositions of moral citizenship. But the notion of 'teacher resilience', whereby a set of values understood as a structure of feeling – following Raymond Williams (1961) – emerged as something that endures across generations and this was partly explained by the importance of 'emotional labour' in teaching (Kirk and Wall, 2011).

Another example of what historical and archival resources have to offer educational researchers was provided in 2001 by Jonathan Rose, whose *Intellectual Life of the British Working Classes* was one of the first examine the reading practices of past generations. There is some evidence for select members of the well-connected, articulate, document-preserving classes, but what of the little recorded majority? Rose employs various kinds of autobiography and memoir written by those from working-class or other modest backgrounds, people who had usually received very little formal schooling, at least until the middle years of the twentieth century. He also makes good use of library records, educational archives, oral histories and Mass Observation, and early social surveys to create a detailed history of working-class reading.

Using image as historical evidence is a relatively new addition in the history of education pioneered in *Silences and Images* (1999), edited by Ian Grosvenor, Martin Lawn and Kate Rousmaniere. Taking representations of teachers as an example, António Nóvoa (2000) analyses his personal archive of over 600 images from published sources that spans two centuries and a range of continents. Nóvoa notes a consistency of themes equating to what he terms a 'grammar of schooling', consisting of secondary teachers undertaking pupil assessment and male primary teachers represented as disciplining figures. In so doing, he combines quantitative analysis and iconography (understood to mean a range of, or system of, types of image used to convey particular meanings) used as a way of 'reading' visual sources (see also Chapter 36 of the present volume).

Christine Wall (2008) drew on careers literature available in University College London's Institute of Education library on open access and material published by the largest teaching union, the National Union of Teachers (NUT), held in the Trades Union Congress (TUC) Library Collections at London Metropolitan University, to study the formation of gendered teacher identities between 1940 and 2000 in Britain. In so doing, Wall paid particular attention to visual representations of teachers on the front covers of the house journal of the NUT, published as *The Schoolmaster and Woman Teachers' Journal* until 1962, when it became *The Teacher*. Her analysis involved, first, a quantitative

'sorting' of images and from this a closer reading of selected images employed to depict the occupation of teaching ensued. The smaller number of images was selected on the basis of a particular iconography: a set of noticeably recurring, thematic compositions conflating the role of woman teacher with motherhood. One series of compositional similarities that stood out, for example, was the standard Christian iconography of the Madonna and Child.

16.6 A worked example: a biographical approach to the history of education

This section reports a worked example of my own practice as a researcher in the history of education. It indicates several key points in planning and 'doing' such research, the personal motivations and commitments of the researcher-as-historian, and the processes involved, focusing on my biographical project on Mary Bridges Adams (née Daltry, 1855–1939), one of twenty-nine women members of the London School Board (LSB). Set up under the 1870 Education Act, school boards were the most advanced democratic bodies of their day. Ratepayers elected them every three years by secret ballot, and women could both vote and stand for office. Multiple voting and the possibility of giving your vote to one candidate favoured the representation of electoral minorities, especially working people and women. For example, the nine women elected in 1879 constituted 18 per cent of all LSB members. Women's numerical representation in the House of Commons did not match this until the 1997 general election, following the Labour Party's adoption of all-women shortlists (1993–6). My narrative exposes some of the trials, tribulations and benefits of historical and documentary research, with the intention of providing guidance and understanding to researchers in the field (see Martin, J., 2013).

The background to the research

From the 1890s, Mary Bridges Adams (she did not hyphenate her surname but others, including her son, did) played an active public role in the British labour movement, even though women did not achieve the vote in general elections until 1918. At a local level, she spent seven years as a member of the LSB, then the largest and most powerful organ of local government in the world. Mary's voice rose out in particular. She was in a minority of one as a socialist woman of working-class social origin. A unionized worker who called herself the 'representative of organized labour', she joined the National Union of Gas Workers and General Labourers because, she said, she was a gas worker on the platform and a general labourer at home. How influential was she?

By 1900, Mary was well-known as a participant within the broader labour movement and as a campaigner for improvements in working-class education. During the First World War, she was in close touch with the European anti-war movement and threw herself into Russian émigré politics. Guiding campaigns in defence of the right of asylum, she had a range of contacts among suffragettes, trade unionists and international socialists. She urged people to fight the abandonment of industrial rights and guarantees, such as the right to strike and restrictions on the use of child labour, to back the unofficial rank-and-file industrial movement on Clydeside and the educational work of the Scottish Marxist John Maclean (1879–1923). Foes thought her an awful woman; friends like George Bernard Shaw (1856–1950) remembered the power of her oratory. The aim of this project was to research and write a significant, original and debate-changing biography to consider the main project of 'making socialists' from the standpoint of gender. Such an assessment has its difficulties but this case outlines the reasons why a study of Mary Bridges Adams is important.

The research 'problem'

Turning to the past means much more than focusing solely upon bureaucrats and politicians who wielded enormous influence in the official central state. It can also involve historical detective work into those places where British women were most influential in the late nineteenth century: local education policy and practice. Mary Bridges Adams was excluded from high politics for the whole of her political career but she did not wilfully hide herself from history. In her lifetime she preserved myriad press cuttings about her activism, and her public utterances stressed her contribution to education and politics as part of the story of British socialism. A truer picture of the past requires an appreciation that British women, like British working men, played an active role in politics in the years before they obtained the national vote. Researchers have so far built up only a partial picture.

In her day, Mary was a national figure in British socialism. Nonetheless, her voice is absent from the established canons of political history, despite inclusion in the *Dictionary of Labour Biography*. The exception to this is Patricia Hollis's *Ladies Elect*, published in 1989. However, her activities do grace the footnotes of some educational histories (Simon, 1965; Kean, 1990; Manton, 2001). Building on this, the objectives of the Bridges Adams project were to: (1) offer a timely and

wide-ranging reappraisal of an era of radical social reform seen through the lens of a pivotal figure; (2) piece together the various parts of Mary's life: public and personal, open and hidden; (3) provide context and a conceptual tool for understanding developments at a crucial turning point in English education, 1890–1910; and (4) challenge political narratives and promote different ways of thinking about the place of the educational question in the study of British socialism.

Mary's life, her attitudes and actions, her role in respect of campaigns for improvements in working-class education, are accessible only through the documentation that has survived. We do not know where she was educated, the schools in which she taught or what kind of classroom teacher she was. Yet she was a prolific writer of articles and her national reputation as a speaker meant that her political activities were regularly reported, and some of her spoken words were relayed in local newspapers. Mindful of the risks of formlessness, the objective was to explore and assess the life's work of Mary Bridges Adams using a range of sources – textual, visual, oral – in ways that allow the reader to understand complexity rather than force Mary's experiences into an over-simplified pattern.

Initiating the process of researching the past: the importance of archives

It was a hot July day in 1990 when I first read Mary's passionate speech in support of free school meals in a dusty copy of the *School Board Chronicle* in the London Metropolitan Archives. Just after her election as a member for Greenwich in the autumn of 1897, there she was at a public Board meeting effectively telling the upper-middle-class membership of the LSB that they could not possibly imagine what it was like to be poor. Writing this woman's life had value for me because of her lifelong concern for class justice. If she had emerged victorious from the political battles that she fought at a crucial turning point in English education, my mother's family would not have felt the sting that came from having to leave school at thirteen or fourteen. She hooked me then and she hooks me now. Fairly quickly, I found myself feeling possessive of my subject rather like the woman researcher A. S. Byatt describes in her brilliant 1990 novel *Possession*.

My starting point was to go to sources: institutional records, personal papers, previously unknown and under-utilized contemporary material, autobiographies, biographies and biographical dictionaries, notably the *Dictionary of Labour Biography*. The *School Board Chronicle* proved crucial because it offered a blow-by-blow account of weekly debates at the meeting of the whole Board. Mary's rhetorical skills were evident, even if she did not always influence decision making. Ahead of her time, she supported the extension of opportunities for all working-class children and specifically attacked the idea of a meritocratic 'ladder of opportunity' for the few.

One of 'my women' at the heart of what started life as a collective investigation of women and educational policy making in late Victorian and Edwardian England (Martin, 1999), Mary was little more than a footnote in the history of education when I began my learning journey. Education politics provided early opportunities for Victorian women and, initially, my main research questions focused on the workings of gender and power on the London School Board, investigating the 'success' factors that facilitated women's careers in public life and the impact of their presence. Mary was one of these pioneer political women, serving from 1897 to 1904 (when the Board was abolished under the 1903 Education Act). When it came to interpreting the sources, meaning could be imputed, not always demonstrated. It is hard to apply key sociological concepts such as power, authority and control to second-hand accounts of a given historical situation. Ambitious plans to employ all sources quickly evaporated given the constraints of time and space. A more realistic decision was taken to focus on the official material on the LSB and the Board's official organ *The School Board Chronicle* that included details of debate at the weekly meetings of school boards throughout England and Wales with extensive coverage of events in London. Other published commentary consulted included the teachers' press, contemporary periodicals and local newspapers, supported by reference to the archives of other relevant institutions, persons and parliamentary papers.

As my gaze turned to the search for Mary's story, I tracked down the references that followed Mary's entry in the *Dictionary of Labour Biography*. Texts were sampled to address specific aspects of my main questions about gender and power: What was the way in which Mary's life unfolded within, and was shaped by, and helped to shape, a particular political, economic, social and cultural context? What were Mary's distinctive qualities – personal and political? What was the impact of her contribution to developments in working-class education, as a committed democrat, socialist and internationalist? What were the dilemmas and contestations that she encountered? What were the criticisms of her?

I constructed my account of Mary's public life from a wider variety of sources, including her own published letters and articles culled from the newspapers of the day, reports in the socialist and educational press,

official records, memoirs and other autobiographical writings, present-day books and articles. I benefited enormously from an oral history interview with Mary's grandson, Nicholas, and the support of his widow, Jenifer. Both offered rare vignettes of a personal kind.

The LSB *Minutes* contain much informative detail, such as data on attendance at meetings, voting records and motions of policy. A search of the index took me to Mary's first appearance via the correspondence pages for 3 May 1895. On that day, six working-class political organizations wrote to request that she inherit the Greenwich seat of a recently deceased Board member. 'Mrs. Bridges Adams, from her learning, great scholastic experience, lucidity of thought and expression, aptness of resource and charm of presence would add distinction to the Board', Woolwich Gas Workers wrote. 'She had the full support, during the last Election of all the Labour Bodies and polled the largest number of votes of any defeated candidate.' In contravention of established protocol, Board members voted to co-opt a male lawyer. Keir Hardie (1856–1915), leader of the UK's Independent Labour Party, angrily protested how the established parties closed ranks to keep a socialist out, with the one dissentient hastily explaining his error (*Labour Leader*, 18 May 1895, p. 8).

Besides the published *Minutes*, background information on the achievements of the LSB was gleaned from the annual report and address of the chairman (available on open access). I was able to order relevant records produced by the sub-committees on which Mary served, together with those on the Board schools for which she became a manager. The amount of information I gathered from the committee records varied. Sometimes I would spend what seemed an eternity hunting down documents that turned out to have nothing much of interest. Sometimes I would use files that were a gold mine of information, such as the 1899 Report on a Special Committee of the General Purposes Committee, particularly helpful on the question of Mary's campaign for school meals, or the 1907 Report on the Bostall Wood Open-air School that ran for three months in the summer of 1907. The experimental open-air school was Mary's initiative, conducted on land owned by the Royal Arsenal Co-operative Society (RACS). The support of the RACS, the largest of its kind in Greater London, was critical to Mary's successful candidature for the LSB. Information about their association can be found in the Half-Yearly Reports and the Bulletin of the Society, known as *Comradeship*.

Tuberculosis was the prime cause of child deaths at the time and doctors claimed they could identify a child at risk, estimated as 10 per cent of the school population. In this context, a new category of 'pre-tuberculous' child became the focus of work on the health benefits of fresh air, sunshine, healthy food and daily exercise. Mary wanted to establish a national system of open-air recovery schools and asked the RACS to donate some of its land at Bostall Wood, Plumstead, for an experiment. The cooperators agreed to her request, the London County Council accepted their offer, and the mixed school opened on 22 July 1907, with 113 weak children. A close reading of press cuttings filed by the cooperators that survive within the RACS Papers in Woolwich, London, shows that 143 newspapers discussed the venture, but only *Clarion*, *Justice*, the *Lancashire Post* and the *Morning Leader* mention Mary's contribution. Her supporters noticed this. Under the headline 'Credit where it is not due', *Justice*, the weekly paper of the Marxist Social Democratic Federation, accused the Conservative councillor Ernest Gray (1857–1932) of trying to gain recognition for a scheme which was not of his making.

Widening the search: uncovering public and private resources and some practical considerations

Much of Mary's evidence is missing because her son destroyed the letters written by her to him, and all photographs of her. As a consequence, her personal papers are only fragments and her surviving letters are mostly short. Widening the search, I was able to track down a mysterious suitcase of papers left with Mary by a leading Russian émigré deported after the Revolution in 1917, now the Mary Bridges-Adams Collection on the British Labour Movement and Russian Socialists, 1905–39, held in the Rare Book and Manuscript Library at Columbia University. Besides Russian letters, they included illuminating correspondence with Mary's patron, the Countess of Warwick (1861–1938).

In association with the study I have compiled a modest database of biographical information to investigate Mary's networking through the use of prosopography, a historical method involving the examination of a number of lives in a given place, to show that socialism was both a lifestyle and a form of organized political activism. This approach facilitated an assessment of social and intellectual backgrounds from a grounded and qualitative perspective, showing the sequencing of connections located in time and space, social history and social geography. All political movements are as much the history of social and intellectual networks as they are of campaigns and lobbying. Making sense of Mary's political career, her circle of influence connected to leftist counter-cultures in Glasgow,

Lancashire, London, the Rhondda and West Yorkshire. Biographical approaches helped generate what Clifford Geertz (1973) called '*thick descriptions*' in ethnography (see Chapter 15 of the present volume). Mary's life is historically placed. The history-writing is a graphic encapsulation of the cultural milieu of which she was a part, to convey a sense of how it felt to be there, the motivations of the struggle, from the point of view of participants who shared her vision of the future.

Accordingly, the works of other historians, political writers and journalists of different hues, biographical writings of fellow travellers and contemporaries were used to contextualize Mary's experience of maturing in the 1870s and 1880s. It is not too difficult to reconstruct some kind of picture of the social structure and life of the nineteenth-century South Wales mining village in which she was born or the Elswick district of Newcastle-upon-Tyne in which she grew up, with the aid of secondary sources, contemporary histories, trade directories, business records, government reports, maps, census material, newspapers, autobiographies and so on. Local history collections were mined to reinsert Mary's presence into the social and political landscape of the period. The best starting point for information relating to local government and politics is the local and regional newspapers for the place in question. In time, knowledge and understanding based on reading the *School Board Chronicle* were supplemented by reports in local newspapers.

My own experience of intermittently researching and writing Mary's biography speaks directly to how the Internet age is affecting scholarship. The expanded capabilities of the web are changing our methods of historical research. Now, digitized guides serve as time-saving devices for the kind of historical detective work described here. Proper integration of a *longue durée* perspective requires me to say that if I were starting the research now it would be far easier to track down the letters produced by Bedford College, for example, that showed how Mary obstinately fought her way into tuition in the Classics.

Up to 2002, I had used the Internet to *prepare* to visit traditional archives, but never to retrieve actual material. So, when I started 'looking for Mary', a local archives service found her in the 1871 Census listed as a fifteen-year-old pupil-teacher (an apprenticeship scheme for training teachers in school settings, common at the time) living with her family at the Robin Adair public house in Elswick. More recently, I was able to use subscriber databases to supplement the family history and the history for the streets in which she lived and those of her political supporters. But it remains the case that the paucity of sources means that I can only speculate where she trained to teach. No school records for the locality have survived (before 1910) and the names of pupil teachers do not appear in the Newcastle School Board *Minutes* until the late 1870s. No clues were found in the Newcastle directories.

Visual remains have been hard to come by, apart from the passport-sized studio photographs of the newly elected London School Board which take up several pages of the *Illustrated London News* in December 1897. Mary is shown in profile. Luxuriant, almost pre-Raphaelite-styled hair frames a strong, attractive face and she fixes the camera with a commanding gaze. This is in stark contrast to her female colleagues, their hair scraped off their foreheads and pulled behind their ears. Ten years later, there is a *Daily Mirror* photograph of two ladies in hats – Mary and the aristocratic Daisy Warwick – resplendent in an open carriage, alongside a report on the 1907 Trades Union Congress. In a staged photo shoot they look directly at the camera with panache. There may be gaps in the account but the real problem in this research centred on how to handle those questions dealing with Mary's attitudes, values, beliefs and aspirations. Writing history 'from the bottom up' meant finding creative ways to scrape that bottom for any smidgen of information.

Building up the political profile

Bridges Adams was among the thousands of women recruited to teach in the nineteenth-century elementary school system. She objected to the unequal social order that she saw strengthened by the educational process, and a vision of a better society carried her out of the classroom and into political action. In her late twenties, she decided to 'cross the river of fire' and enter the socialist movement. This was how William Morris (1834–96) characterized the life-changing experience that becoming a socialist represented in the 1880s and 1890s. Eager idealists, borne along by an almost millennial fervour, the men and women (like Mary) who made up the 'pioneering generation' of British socialists spent much time and energy spreading the word.

Creating the National Labour Education League in 1901 represented a transition in Mary's political journey, putting her on a trajectory of moving between local and national activities at several levels. Against a background of mounting tensions in the educational world, the League heralded a new step forward – the closest thing to a Labour education policy then in existence. For her, the advance of education and political progress were part of a single programme. The abolition of the School Board in 1904 did not diminish an extra-parliamentary activism which focused on

education but couched the question within wider questions of social justice and other interests. With the financial support of her aristocratic patron, Mary used her training in street-corner meetings and outdoor agitation to fight for open-air schools for malnourished and tubercular children, free school meals and medical inspection.

After 1908, Mary was closely involved with the Marxist educators of the Plebs League and the labour colleges. Like them, her conception of education for the workers was rooted in the Marxist studies promoted by the earlier Socialist League slogan: 'Educate, agitate, organise!' She advocated direct action mounted by rank-and-file agitators and Marxist pedagogues. This included support for a strongly political conception of adult education, identifying a distinctively socialist or working-class curriculum. Therefore, she opposed the liberal education philosophy espoused by the Workers Educational Association (WEA, founded in 1903). She was tireless in her efforts to win trade union backing for the principle of working-class self-organization and even secured funding to provide an educational space for politically active working-class women akin to the School of Social Science maintained by the German Social Democratic Party. In 1912, she opened Bebel House in Lexham Gardens, round the corner from the Central Labour College in London's Earls Court, and installed herself as resident principal. In these years she built alliances with radical suffragists, attested to by her attempts to reach working-class Lancashire women as political editor of the *Cotton Factory Times*. Almost every week for many years, the *Cotton Factory Times* included an editorial by her.

Recently opened Home Office files on some of the Russian émigrés with whom she worked helped reconstruct Mary's activities in the First World War, shedding light on her character and personality (though individual documents remained closed). Press cuttings kept by her opponents in the WEA were deeply revealing and the librarian at the TUC Library Collections kindly brought them to my attention when they were deposited there. Some of Mary's surviving letters were excavated from the Sheehy Skeffington Papers held in the National Library of Ireland, which include Mary's letter of sympathy concerning the murder of Francis Sheehy Skeffington in Dublin at Easter in 1916 and active support for his widow Hanna's campaign for a public inquiry.

Closing the loop

Mary's activities are useful for plotting women's roles in British leftist 'oppositional networks'. The experience was a contradictory one, but being a woman was only part of this. There was also her preparedness to challenge orthodoxy, demonstrated by her writings and action, all of which suggests a deep commitment to furthering the cause of socialism in a time of intense conflict over the shape and purpose of education. We know she troubled the establishment elite, but in assessing her contributions to the conflicts of the years 1890 to 1910, we need to question the conventional wisdoms since, more often than not, they vindicated the wisdom of the powers that be. In so doing we need to go back to the margins, to listen to the testimony of those to whom she was closest. Mary's specific utility as a historical subject is to represent the common, unnamed socialist woman who fought to bring class consciousness into being at a specific conjuncture. For Mary, education was the path to a new social order.

16.7 Conclusion

To those who would do the spade work, different approaches are possible even with the existing sources. It *is* possible to recreate 'lost lives' and events by going beyond the official record and digging for the raw material of history. There is a remarkable amount of unexploited personal and ordinary information out there and digitization is opening up all kinds of possibilities. Ultimately, there is great joy and sense of satisfaction to be had when a historical 'puzzle' seemingly falls into place. Despite the attrition of memories and histories I *did* manage to tell Mary's story and thereby reveal a valuable history long overlooked. I was uplifted when, in closing the hermeneutic circle, a colleague commended my journey through social histories, biography and politics (Martin, 2010) as elucidating the fictional past contrived in A. S. Byatt's *The Children's Book* published in 2009: embedded in the thoughts, beliefs and feelings of late-Victorian and Edwardian England.

But whatever the category of evidence, the past does not in any automatic way 'speak for itself'. It *can* prove impossible to locate information we 'expect' to find, so be prepared for frustrations and disappointments and allow your research questions to be guided by the material available, rather than establishing an experimental hypothesis approach. Outstanding history-writing involves conveying to readers something of the processes by which the raw materials of history have been produced.

To return to the notion of a 'foreign country' with which our narrative began: they *did* do things differently there, but we owe a duty to past, present and future generations to represent *all* those pasts and *all* those interlocutors as accurately as possible (see also

Lowenthal, 2015). When it comes to our educational past, historical and documentary research needs to include not just the leaders but the people who inhabit the classrooms and the forms, dimensions and meaning of their experience, to bring history into and out of the community and ensure that the unknown ideas from the underclasses, the unprivileged and the defeated, are told with integrity and not quietly forgotten. In this way, a theoretically informed history of education that gives recognition to alternative trajectories and the road not followed can make a fine contribution to the work of creating a better future.

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Surveys, longitudinal, cross-sectional and trend studies

CHAPTER 17

There are many different kinds of survey; each has its own characteristics and key issues. We set these out in this chapter, addressing such matters as:

- what is a survey?
- some preliminary considerations
- planning a survey
- low response and non-response, and how to reduce them
- survey sampling
- longitudinal, cross-sectional and trend studies
- strengths and weaknesses of longitudinal, cohort and cross-sectional studies
- postal, interview and telephone surveys
- comparing methods of data collection in surveys

17.1 Introduction

We advise readers to take this chapter in conjunction with Chapters 12 (sampling), 24 (questionnaires), 25 (interviews) and data-analysis techniques (Part 5). Many researchers reading this book will probably be studying for higher degrees within a fixed and maybe short time frame; that may render longitudinal study out of the question for them. Nevertheless longitudinal study is an important type of research, and we introduce it here. More likely, researchers for higher degrees will find cross-sectional survey research appropriate, and it is widely used in higher degree research.

In many quarters, Internet surveys are becoming the predominant method of surveys, through email (with a questionnaire as an attachment, or embedded in the email, or with a hyperlink link to a website, social networking site, special interest group, listserv, discussion group etc.), with companies providing free or low-cost software to design questionnaires and, indeed, to conduct the survey and collect data for researchers. Given the rise and widespread usage of Internet surveys, we devote an entire, separate chapter (Chapter 18) to this. However, we include reference to Internet surveys in Table 17.3 in this chapter, for purposes of comparison with other means of survey design and conduct.

17.2 What is a survey?

Many educational research methods are descriptive; that is, they set out to describe and to interpret what is. Such studies look at individuals, groups, institutions, methods and materials in order to describe, compare, contrast, classify, analyse and interpret the entities and the events that constitute their various fields of enquiry. We deal here with several types of survey research, including longitudinal, cross-sectional and trend or prediction studies.

Typically, surveys gather data at a particular point in time with the intention of describing the nature of existing conditions, or identifying standards against which existing conditions can be compared, or determining the relationships that exist between specific events. They may vary in their levels of complexity, from those which provide simple frequency counts to those which present relational analysis.

Surveys may be further differentiated in terms of their scope and complexity. A study of contemporary developments in post-secondary education, for example, might encompass the whole of Europe; a study of subject choice, on the other hand, might be confined to one secondary school.

17.3 Advantages of surveys

A survey has several characteristics and several claimed attractions; typically it is used to scan a wide field of issues, populations, programmes, people etc. in order to measure or describe any generalized features. It is useful (OECD, 2012; Dillman *et al*., 2014) in that it often:

- gathers data on a one-shot basis and hence is economical and efficient;
- represents a wide target population (hence there is a need for careful sampling, see Chapter 12);
- generates numerical data;
- provides descriptive, inferential and explanatory information;
- manipulates key factors and variables to derive frequencies (e.g. the numbers registering a particular opinion or test score);

- gathers standardized information (i.e. using the same instruments and questions for all participants);
- ascertains correlations (e.g. to find out if there is any relationship between gender and test scores);
- presents material which is uncluttered by specific contextual factors;
- captures data from multiple-choice, closed questions, test scores or observation schedules;
- supports or refutes hypotheses about the target population;
- generates accurate instruments through piloting and revision;
- makes generalizations about, and observes patterns of response in, the targets of focus;
- gathers data which can be processed statistically;
- uses large-scale data gathered from a wide population in order to enable generalizations to be made about given factors or variables.

Examples of surveys are:

- test scores (e.g. from students nationally, internationally, locally);
- students' preferences for particular courses, for example, humanities, sciences;
- attitudes to, and opinions of, quality of teaching;
- surveys of groups of people's values over time;
- surveys of factors (e.g. income levels, social class membership, inequality) over time;
- opinion polls;
- reading and mathematics performance surveys.

Surveys in education often use test results, self-completion questionnaires and attitude scales. Here a researcher may be seeking to gather large-scale data from as representative a sample as possible in order to say with a measure of statistical confidence that certain observed characteristics occur with a degree of regularity, or that certain factors cluster together (see Chapter 43) or that they correlate with each other (correlation and covariance), or that they change over time and location (e.g. results of test scores used to ascertain the 'value-added' dimension of education, maybe using regression analysis and analysis of residuals to determine the difference between a predicted and an observed score), or regression analysis to use data from one variable to predict an outcome on another variable.

Surveys can be *exploratory*, in which no assumptions or models are postulated, and in which relationships and patterns are explored (e.g. through correlation, regression, stepwise regression and factor analysis). They can also be *confirmatory*, in which a model, causal relationship or hypothesis is tested (see the discussion of exploratory and confirmatory analysis in Chapter 43). Surveys can be descriptive or analytic (e.g. to examine relationships). Descriptive surveys simply describe data on variables of interest, whilst analytic surveys operate with hypothesized predictor or explanatory variables that are tested for their influence on dependent variables or relationships between variables.

Many surveys combine nominal data on participants' backgrounds and relevant personal details with other data (e.g. attitude scales, data from ordinal, interval and ratio measures) (see Chapter 38). Surveys are useful for gathering factual information, data on attitudes and preferences, beliefs and predictions, opinions, behaviour and experiences – both past and present (Weisberg *et al*., 1996; Aldridge and Levine, 2001; Dillman *et al*., 2014). Their attraction lies in their appeal to generalizability or universality within given parameters, their ability to make statements which are supported by large data and their ability to establish the degree of confidence which can be placed in a set of findings.

On the other hand, if a researcher is concerned to catch local, institutional or small-scale factors and variables – to portray the specificity of a situation, its uniqueness and particular complexity, its interpersonal dynamics, and to provide explanations of why a situation occurred or why a person or group of people returned a particular set of results or behaved in a particular way in a situation, or how a programme changes and develops over time – then a survey approach may be unsuitable. Its explanatory potential or fine detail is limited; it is lost to broad-brush, often descriptive generalizations which are free of temporal, spatial or local contexts. Williams *et al*. (2016) note that having a two-phase process of a postal survey – an initial screening survey followed by the topic-based survey sent to eligible people – is also a useful device for obtaining more in-depth data. In a survey the individual instance is sacrificed to the aggregated response (which has the attraction of anonymity, non-traceability and confidentiality for respondents and opportunity for trends and patterns to be discovered).

Surveys typically, though by no means exclusively, rely on large-scale data, for example, from questionnaires, test scores, attendance rates, results of public examinations etc., all of which enable comparisons to be made over time and between groups. This is not to say that surveys cannot be undertaken on a small-scale basis, as indeed they can; rather it is to say that the generalizability of such small-scale data will be slight. In surveys the researcher is usually an outsider; indeed questions of reliability and possible bias can attach

themselves to researchers conducting survey research on their own subjects, for example, participants in a course that they have been running. Further, it is critical that attention is paid to rigorous sampling, otherwise the basis of the survey's applicability to wider contexts is seriously undermined. Non-probability samples tend to be avoided in surveys if generalizability is sought; probability sampling will tend to lead to generalizability of the data collected.

17.4 Some preliminary considerations

A fundamental decision by the researcher is whether a survey is the appropriate means of answering the research purposes and research questions (Magee *et al*., 2013). Assuming that it is, three prerequisites to the design of any survey are: the specification of the exact purpose of the enquiry; the population and issues on which it is to focus; and the resources that are available. Hoinville and Jowell's (1978) consideration of each of these key factors in survey planning can be illustrated in relation to the design of an educational enquiry.

The purpose of the enquiry

First, a survey's general purpose must be translated into a specific central aim. Thus, 'to explore teachers' views about in-service work' is somewhat nebulous, whereas 'to obtain a detailed description of primary and secondary teachers' priorities in the provision of in-service education courses' is reasonably specific.

Having decided upon and specified the primary objective of the survey, the second phase of the planning involves the identification and itemizing of research questions which will enable the objective to be addressed. The third phase, usually driven by the research questions, is to identify subsidiary topics that relate to its central purpose. In our example, subsidiary issues might well include: the types of courses required; the content of courses; the location of courses; the timing of courses; the design of courses; and the financing of courses.

The fourth phase follows the identification and itemization of subsidiary topics and involves formulating specific information requirements relating to each of these issues. For example, with respect to the type of courses required, detailed information would be needed about the duration of courses (one meeting, several meetings, a week, a month, a term or a year), the status of courses (non-award bearing, award bearing, with certificate, diploma, degree granted by college or university), the orientation of courses (theoretically oriented involving lectures, readings, etc., or practically oriented involving workshops and the production of curriculum materials).

As these details unfold, consideration has to be given to the most appropriate ways of collecting items of information (interviews with selected teachers, postal questionnaires to selected schools, online questionnaires etc.).

The population upon which the survey is focused

The second prerequisite to survey design, the specification of the population (e.g. people, issues) to which the enquiry is addressed, affects decisions that researchers must make both about sampling and resources. In our hypothetical survey of in-service requirements, for example, we might specify the population as 'those primary and secondary teachers employed in schools within a thirty-mile radius of Loughborough University'. In this case, the population is readily identifiable and, given sufficient resources to contact every member of the designated group, sampling decisions do not arise.

Things are rarely so straightforward, however. Often the criteria by which populations are specified ('severely challenged', 'under-achievers', 'intending teachers' or 'highly anxious') are difficult to operationalize. Populations, moreover, vary considerably in their accessibility; students and student teachers are relatively easy to survey, travellers' children and headteachers are more elusive. More importantly, in a large survey, researchers usually draw a sample from the population to be studied; rarely do they attempt to contact every member. We deal with the question of sampling shortly.

The resources available

Resources are not simply financial. For example, survey design can be costly in terms of time, and consideration of resources has to include human, material, financial, administrative, temporal, geographical, technical (e.g. computer-related) costs. An important factor in designing and planning a survey is financial cost. Sample surveys are labour-intensive, the largest single expenditure being fieldwork, where costs arise out of interviewing time, travel time and transport costs of the interviewers themselves. There are additional demands on the survey budget. Training and supervising the panel of interviewers can often be as expensive as the costs incurred during the time that they actually spend in the field. Questionnaire construction, piloting, printing, posting, coding, together with computer programming and processing all eat into financial resources.

Mode of data collection

There are two main issues to be addressed here:

1. Will the researcher be completing the survey by entering data, or will the participants be self-administering the survey?
2. How will the survey be administered, for example, a postal survey, a telephone survey, an Internet survey, by face-to-face interviews, group-administered surveys, self-administered surveys, drop-off surveys, email? A full account of the interview as a research technique is given in Chapter 25.

Dillman *et al*. (2014) advise researchers to use multiple and mixed modes of delivery/administration, as this helps response rates.

Self-reporting

There can be a large difference in the responses gained from self-reporting and those obtained from face-to-face survey interviews or telephone interviews (Dale, 2006, p. 145; Dillman *et al*., 2014). Many surveys ask respondents not only to administer the questionnaires themselves but also to report on themselves. This may introduce bias, as respondents may under-report (e.g. to avoid socially undesirable responses) or over-report (to give socially desirable answers). Self-reporting also requires the researcher to ensure that: respondents all understand the question, understand it in the same way and understand it in the way intended by the researcher (Kenett, 2006, p. 406). The difficulty here is that words are inherently ambiguous (see Chapter 24 on questionnaire design), so the researcher should be as specific as possible. The researcher should also indicate how much contextual information the respondent should provide, what kind of answer is being sought (so that the respondent knows how to respond appropriately), how much factual detail is required and what constitutes relevant and irrelevant data (e.g. the level of detail or focus on priority issues required) (pp. 407–8). Further, surveys that rely on respondents' memory may be prone to the bias of forgetting or selective recall.

Ethics

Ethical issues are discussed in Chapters 7, 8 and 18, and we refer readers to these; here we note the importance of gaining the informed consent of respondents. Whilst completion of the survey might be taken as giving consent, this may not always be the case, and the completion of a consent form may be needed (though some participants may be suspicious of this), and indeed Dillman *et al*. (2014) note that asking for consent requires the researcher to make it clear what the consent is being given for, as, for example, to ask for consent before the questions have been asked is asking participants to take a leap of faith. We also note in Chapters 7 and 28 that informed consent is complex, as it is unclear what is being consented to, and for how long, and for what purposes and uses, and that these problems are exacerbated when data are archived for future use as secondary data sets. Informed consent should also include the right not to participate or to withdraw at any time.

Ethical issues here also concern attention to confidentiality, anonymity, privacy and non-traceability. In paper-based surveys this may be easy to guarantee, but, as we indicate in Chapter 18, for electronic and Internet-based (e.g. website and email surveys), no such absolute guarantees are available. Such computer-related problems raise the matter of data security and identity protection. In electronic and paper surveys, telephone interviewing and face-to-face surveys, the researcher might not ask for, or require, identifying features, or might remove these when storing and archiving data.

However, in group interviews these may not be so easy to protect (e.g. members of the group may talk to others), and in electronic/Internet-based surveys, the service provider can log and track participants, and data miners and hackers can break into data, particularly email, even when security steps have been taken. We discuss this in Chapter 18.

The researcher can, and should, take all reasonable steps to protect confidentiality, anonymity, privacy and non-traceability and indicate to respondents what those steps are, recognizing that where there are limits (e.g. in electronic surveys), this may lead to some respondents not taking part.

Underpinning ethical issues in surveys is the requirement of *primum non nocere*: primarily, do no harm. The researcher must take every step necessary to address this. This concerns access to, collection, storage, use, dissemination and reporting of data, and subsequent archiving of data or locating the data in the public domain, with immense care being taken with regard to identification and sensitive information. This raises issues not only of removing identifying features, removing certain data, aggregating or anonymizing data, but who owns the data and what rights the owner has, once the data have been given to the researcher. The researcher has a duty of care and of trust here.

17.5 Planning and designing a survey

Whether the survey is large scale and undertaken by some governmental bureau, or small scale and carried

out by the lone researcher, the collection of information typically involves one or more of the following data-gathering techniques: structured or semi-structured interviews, self-completion (e.g. postal and Internet questionnaires), telephone interviews, Internet surveys, standardized tests of attainment or performance, and attitude scales.

Planning a survey involves knowing: (a) what exactly you wish to find out, and why; (b) what data you need to be able to answer (a); (c) what questions you will ask to acquire the data. Researchers must also consider: sample selection and access to the sample; distribution/data collection and return of surveys; measurement design and data types; ethical issues; piloting; analysis and reporting.

Sapsford (1999, pp. 34–40) suggests that there are four main considerations in planning a survey:

- *problem definition* (e.g. deciding what kinds and contents of answers are required; what hypotheses there are to be tested; what variables there are to explore);
- *sample selection* (e.g. what is the target population; how can access and representativeness be assured; what other samples will need to be drawn for the purpose of comparison);
- *design of measurements* (e.g. what will be measured, and how (i.e. what metrics will be used – see Chapter 24 on questionnaires); what variables will be required; how reliability and validity will be assured);
- *concern for participants* (e.g. protection of confidentiality and anonymity; avoidance of pain to the respondents; avoiding harm to those who might be affected by the results; avoiding over-intrusive questions; avoiding coercion; informed consent; see Chapters 7 and 8).

Typically surveys proceed through well-defined stages, outlined in Figure 17.1. Though these are set in a

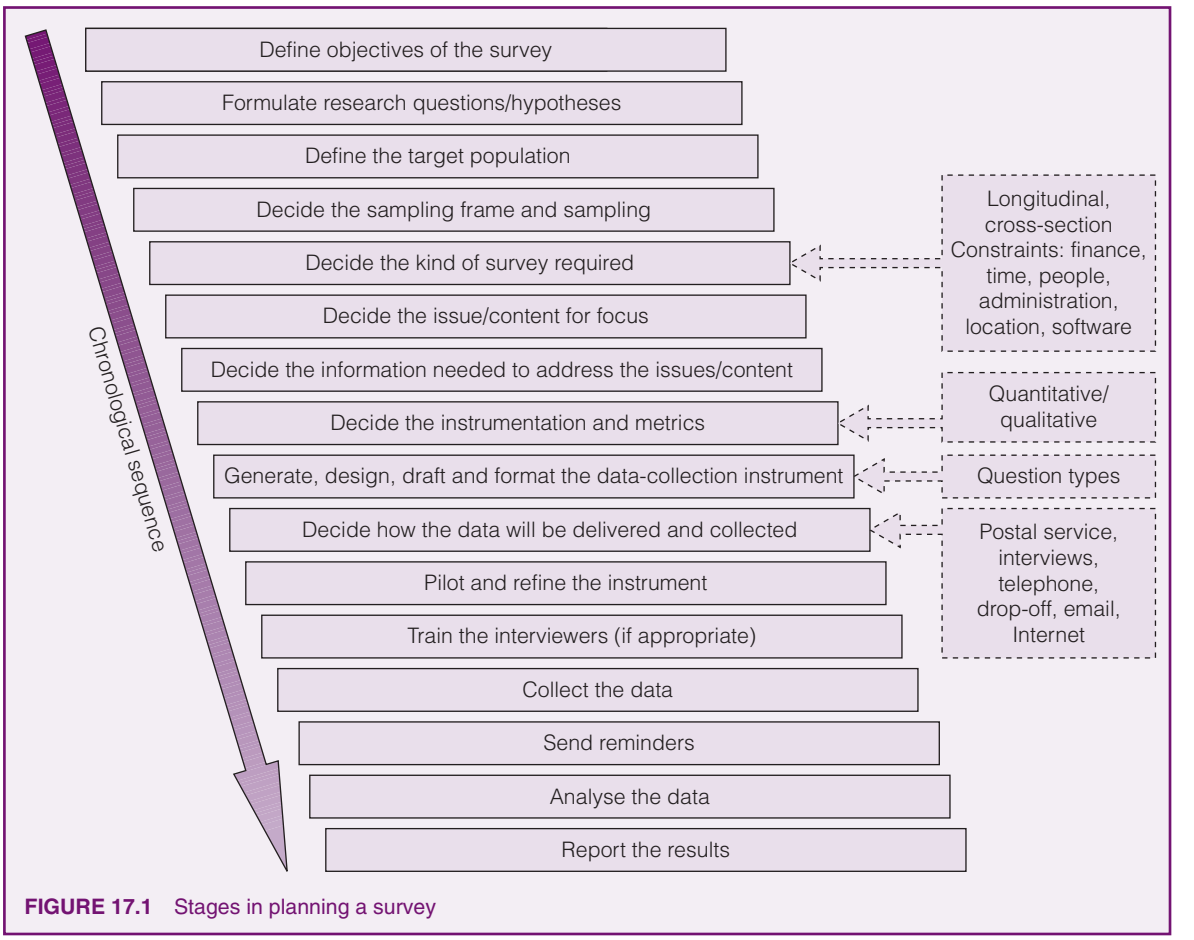


FIGURE 17.1 Stages in planning a survey

sequence, the sequence may alter and the process is iterative and recursive. The process moves from the general to the specific. A general research topic is operationalized into component issues and questions, and, for each component, questions are set. As with questionnaires (Chapter 24), it is important, in the interests of reliability and validity, to have several items or questions for each component issue, as this does justice to the all-round nature of the topic.

Rosier (1997, pp. 154–62) suggests that the planning of a survey must include clarification of:

- the research questions to which answers need to be provided;
- the conceptual framework of the survey, specifying in precise terms the concepts that will be used and explored;
- operationalizing the research questions (e.g. into hypotheses);
- the instruments to be used for data collection, for example, to chart or measure background characteristics of the sample (often nominal data), academic achievements (e.g. examination results, degrees awarded), attitudes and opinions (often using ordinal data from rating scales) and behaviour (using observational techniques);
- sampling strategies and sub-groups within the sample (unless the whole population is being surveyed, e.g. through census returns or nationally aggregated test scores etc.);
- pre-piloting the survey (to generate items for the survey);
- piloting the survey;
- data-collection practicalities and conduct (e.g. permissions, funding, ethical considerations, response rates);
- data preparation (e.g. coding, data entry for computer analysis, checking and verification);
- data analysis (e.g. statistical processes, construction of variables and factor analysis, inferential statistics);
- reporting the findings (answering the research questions).

Ruel *et al*. (2015) comment that researchers need to consider:

- the kind of survey to be used;
- ethical issues;
- questionnaire and instrument design and appearance;
- question construction (measures, responses and measurement error);
- validity and reliability;
- sampling;
- response rates, non-responses and attrition;
- the medium of delivery, completion and return of the survey;
- data entry and data cleaning;
- data analysis and reporting;
- missing data;
- data archiving.

It is important to pilot and pre-pilot a survey. The difference between the pre-pilot and the pilot is this: the pre-pilot is usually a series of open-ended questions that are used to generate items and categories for closed, typically multiple-choice questions, whilst the pilot is used to test the draft of the actual survey instrument itself (see Chapter 24).

A rigorous survey formulates clear, specific objectives and research questions; ensures that the instrumentation, sampling and data types are appropriate to yield answers to the research questions; and ensures that as high a level of sophistication of data analysis required can be done (i.e. as the data will sustain).

Attention must be given to: the mode of data collection; respondent effort (too much and this can lead to non-response); question wording, sequence and format.

Some challenges in planning surveys

A survey is no stronger than its weakest point, and we consider a range of issues here in order to strengthen each aspect of a survey (e.g. OECD, 2012). Surveys must minimize errors caused by:

- poor sampling (e.g. failure to represent or include sufficiently the target population);
- poor question design and wording (e.g. failure to catch accurately the views of, or meanings from, the respondents or to measure the factors of interest);
- incorrect or biased responses;
- low response or non-response.

The first of these – a sampling matter – may be caused by a failure correctly to identify the population and its characteristics, or a failure to use the correct sampling strategy, or systematically to bias the sample (e.g. using a telephone survey based on telephone directory entries, when key people in the population – the poor – may not have a telephone, or may have a cellphone rather than a fixed line (the young, the middle aged but not the elderly), or using an Internet- or email-based survey when many respondents do not have access). We address sampling issues in Chapter 12 and below.

The second of these is a failure to operationalize the variables fairly (i.e. a validity issue) or a failure in the

wording or meanings used or inferred, such that incorrect responses are collected (a reliability issue) (e.g. people may not understand a question, or may misinterpret it, or interpret it differently). We address this in Chapter 14 and below.

The third problem is that some participants may deliberately over-report or under-report the real situation in – often sensitive – matters. For example, teenage alcohol, smoking or drug use, underage sexual relations, bullying, domestic violence, petty criminality may be *systematically* under-reported (i.e. be biased), whereas the popularity of a teacher or students might be over-reported (i.e. biased). Bias obtains where there is a *systematic* skewing or distortion in the responses.

Further, some questions may rely on memory, and memory can be selective and deceptive (e.g. people may not remember accurately). Also, some responses will depend on a person's state of mind at the time of completing the survey – asking a teacher about teacher stress and tiredness late on a Friday afternoon in school with a difficult class could well elicit a completely different response from asking her directly after a week's holiday. Some questions may be so general as to be unhelpful (e.g. 'how stressed do you feel?'), whereas others might be so specific as to prevent accurate recall (e.g. 'how many times have you shouted at a class of children in the past week?') (one solution to the latter might be to ask participants to keep a diary of instances).

Fowler (2009, p. 15) suggests that a respondent's answer is a combination of the true response plus an error in the answer given, with errors coming from many sources.

The fourth of these – low response or non-response – is a problem that besets researchers, and is so significant that we devote a separate section to it below.

Dillman *et al*. (2014) identify four key errors to be avoided in surveys which seek to represent a wider population:

- coverage error (poor and incomplete representation of the population in the sample). For example, a coverage error might be made if telephone or Internet surveys are used, as not everyone has a telephone (particularly a landline) or access to, and familiarity with, the Internet;
- sampling error (including inaccurate estimates of the population);
- non-response error (the difference between a representative result and that obtained from non-response of different individual or groups, i.e. a skewed response); and
- measurement error: inaccurate and unreliable response because of (a) the metrics, scales and units of measurement used; (b) socially desirable responses and respondent acquiescence (the tendency to agree with an interviewer rather than disagree) in face-to-face survey interviews; (c) questionnaire features, for example, length, difficulty, questions asked, complexity, order effects, interviewer effects, survey mode (post, telephone, email, interview, Internet etc.).

17.6 Survey questions

Though we go into detail about questions and questionnaires in Chapter 24, here we give advice on some important issues in writing and asking questions in surveys (Creswell, 2012; OECD, 2012; Abascal and Diaz de Rada, 2014; Champagne, 2014; Dillman *et al*., 2014; Colorado State University, 2016):

- Ensure that the questions cover the topics and research questions comprehensively and with the appropriate scales of measurement and scales (e.g. 1–5, –4 to +4, 'strongly disagree' to 'agree').
- Keep the survey simple and short, and use whole, short sentences.
- Consider respondent effort: avoid overloading the respondent with thinking, recalling, reading and responding.
- Ensure that the questions apply to all the respondents.
- Consider the order of the questions (questions are not independent of each other, and the answer to one question may affect the answer to another in the respondent's mind, e.g. the primacy effect, 'carry over' and 'anchoring effect' (Dillman *et al*., 2014, p. 235), i.e. what comes first affects what comes later and respondents use the early questions as a standard against which they compare the later questions).
- Arrange the order and organization of the survey in a way that is easy for the respondent to understand (subheadings in a written survey are important here).
- Group together questions that cover similar topics, with subheadings in written surveys, to parallel what would naturally happen in a conversation (NB if respondents see two questions as similar then, for consistency, they will give answers which are similar).
- Start the survey with questions that respondents will find meaningful and interesting, and will be able to answer.

- If you are using branching questions, ask all the branching questions before you ask the follow-up questions.
- Ensure that the wording is comprehensible to the respondent (use easy words) and judge how the respondent will regard and feel about the question asked.
- Keep sensitive questions until later in the survey.
- Avoid putting the important questions right at the end of the survey.
- Consider the willingness of the respondent to answer the questions correctly and honestly, and whether the respondent will actually know the answer (e.g. to factual questions or to questions which require long-term memory), i.e. whether the question really applies to the respondent.
- Consider what the question is asking for – for example, factual answers; attitudes, perceptions and opinions; behaviours; events – and how to make these clear to the respondent. Some factual information is easy (e.g. gender, age) but other data (e.g. attitudes, behaviours, those which rely on memory) may be less accurate.
- Use concrete, specific and precise terms (define terms concretely) so that the respondent understands exactly what is being asked for in the survey.
- Consider the suitability of question types and formats: (a) for *nominal variables*: dichotomous, multiple choice (single choice, restricted number of choices, free number of choices); (b) for *ordinal variables*: rating scales, ranking scales; (c) for *interval, ratio and continuous variables*: constant sum, percentages/marks out of ten, open number (e.g. number of hours of study in a week); (d) for *non-numerical answers*: open questions. Decide whether to have a mid-point in scale items; use large-range scales if subsequent factor analysis is intended; and ensure that response categories are exhaustive, to fit the choices that participants will really want, i.e. that they enable respondents to say what they want to say (and this underlines the importance of running a pilot).
- Avoid: double-barrelled questions (asking more than one thing in a single question); long and complex questions and vocabulary; technical language; negatively worded items; ambiguous questions; leading questions (those which influence the response and indicate a desired response); questions which may cause embarrassment.
- Consider the medium of the administration/conduct/'delivery' of the survey, for example, postal service, email, face-to-face interview, website, telephone, i.e. visual, oral and aural administration of the survey, and who enters the responses (the respondent or the interviewer).
- Consider whether it is advisable to have an interviewer present or absent, as the interviewer's presence may bias the respondent, raising issues of the respondent's concern for (a) social desirability and (b) acquiescence (defined above); acquiescence is a particular problem in questions which include 'agree', as there is a tendency to agree.

Magee *et al*. (2013) advise researchers to consider:

- how others have addressed the constructs in question; developing and writing relevant survey items clearly;
- the mode of the item, for example, a statement or a question (a question is preferable);
- the response (number and type, with no smaller than a five-point scale; odd numbers or even numbers in scaling; inclusion of positive and negative options or only positive options: avoid agreement- or positive-only responses; label each point in an ordinal scale);
- reliability and validity of items;
- ensuring that the question is interpreted by respondents in the way intended.

Given these points, it is essential that a survey be piloted, and we give guidelines to piloting in Chapter 24, for example, for content, coverage, ease of understanding, timing, redundancy, sensitivity, question types, question order, mode of delivery, ease of completion, answerability.

17.7 Low response, non-response and missing data

Response and non-response are related to contact, cooperation and ease of conduct, completion and return of the survey (Dillman *et al*., 2014). Non-response to a whole questionnaire ('unit non-response'; Durrant, 2009, p. 293) or to a specific item ('item non-response'; p. 293) is a serious problem for much survey research, though Denscombe (2009b, p. 282) notes that online surveys tend to have lower item non-response than paper-based surveys, though there may be more drop-outs before reaching the end of an online survey than in a paper-based survey.

Dale (2006, p. 148) suggests that 'non-respondents almost invariably differ from respondents', and that this affects the validity and reliability of the responses obtained, and their analysis. If non-response is received from a very homogeneous sample then this might be

less of a problem than if the sample is very varied. Further, if non-response is received randomly across a sample then this might be less of a problem than if the non-response was from a particular sub-sector of the respondents (e.g. a very low or a very high socio-economic group), as this would bias the results (cf. Dale, 2006, p. 148). A subset of non-response to a whole questionnaire is item non-response, and here missing data should not be ignored (Dale, 2006, p. 15).

Rubin (1987), Little and Rubin (1989), Allison (2001), Dale (2006, pp. 149–50) and Durrant (2006, 2009) review a range of different 'imputation methods' for handling and weighting non-response, i.e. methods for filling in missing data with 'plausible values' in order to render a set of data complete and yet to reduce bias in the non-responses, i.e. that bias which might be caused by the non-responses having different values from the non-missing responses (Durrant, 2009, p. 295). These depend on whether the non-response is largely confined to a single variable or many variables. The researcher has to determine whether there are patterns of non-response, as these affect the method for handling non-response. For example, if the non-response is randomly distributed across several variables, with no clear patterns of non-response, then this may be less problematic than if there is a systematic non-response to one or more variables in a survey (Durrant, 2009, p. 295; Dillman *et al*., 2014). Durrant (2009) sets out several ways of calculating missing values, including:

- calculating missing values from regression techniques using auxiliary variables (p. 296);
- 'hot deck' methods, in which sub-groups of participants (based on their scores on auxiliary variables) are constructed and the researcher compares their results to the non-missing results of the respondent who had omitted a particular response (p. 297);
- 'nearest neighbour' techniques, in which the results from a person whose data diverge as little as possible from those of the missing person are used to replace the missing values.

Durrant (2006, 2009) and Dillman *et al*. (2014) identify further, statistical methods of calculating missing scores, such as multiple and fractional imputation and propensity score weighting. Durrant makes the point that how one calculates the values of missing data depends on a range of factors such as the purpose of the analysis, the variable(s) in question, the kinds of data, any patterns of missing data, and the characteristics and fittingness of the assumptions on which the particular intended imputation method is based. The National Centre for Research Methods (2016) also suggests that using means of groups and sub-groups responding to a particular item can be used for imputation. Here one looks for patterns of missing data (any groups of units/cases or items) and calculates an average value (e.g. on a scaled item) for groups/sub-groups of cases (individuals), and reporting standard error.

Ary *et al*. (2002) note that non-respondents may be similar to late responders, so it might be possible to use data from late responders to indicate the possible responses from non-respondents. This requires the researcher to identify late responders.

Missing data within a survey can have many causes. For example, people may not be present on the day of its administration, or they may not understand the question, or they may take exception to the question or overlook it by mistake. Pampaka *et al*. (2016) give the example of the administration of a school survey on bullying, where students may be absent without predictable reasons, or they are representing their school in a competition (e.g. high-performing and highly motivated students), or they may be more likely to be bullied (p. 19). All these, the authors note, lead to biased data. They note that missing data are a particular problem in longitudinal surveys and surveys across phase transitions. They note that statistical analysis (e.g. stepwise regression, which ignores missing data) is dangerous if there are missing data, and they argue for multiple imputation methods. However, they also note that multiple imputation methods are essentially speculative, based on simulations (p. 21).

Pampaka *et al*. (2016) distinguish between missing data from units (individuals) and items, but both can lead to a biased response. There are many ways to address this, for example, by simply analysing incomplete data, or by weighting, and by imputation. Weighting is designed to ensure a better representation of the population, and it can be used to adjust data for non-response, to bring the data into the correct matching of the population. If the incomplete data are random, i.e. all cases have equal probability of being missing (as in their example of those students who are absent for unpredictable reasons), then the analysis may be unbiased (the claim of randomness for equality of distributions, see Chapter 20 on experiments).

For further guidance on weighting, standard error and imputation, we refer the reader to the sources indicated above and to the guidance from the National Centre for Research Methods (www.restore.ac.uk).

In some cases (e.g. when all the students in a class complete a questionnaire during a lesson) the response rate may be very high, but in other circumstances the response rate may be very low or zero, either for the

whole survey or for individual items within it, for several reasons, for example:

- the survey never reaches the intended people;
- people refuse to answer;
- people may not be available (e.g. for a survey administered by interview), for example, they may be out at work when a telephone survey administrator calls;
- people may not be able to answer the questions (e.g. language, reading, speaking or writing difficulties);
- people may not actually have the information requested;
- people may overlook some items in error;
- the survey was completed and posted but failed to return;
- the pressure of competing activities on the time of the respondent;
- potential embarrassment at their own ignorance if respondents feel unable to answer a question;
- ignorance of the topic/no background in the topic;
- dislike of the contents or subject matter of the interview;
- fear of possible consequences of the survey to himself/herself or others;
- lack of clarity in the instructions;
- fear or dislike of being interviewed (or of the interviewer);
- sensitivity of the topic, or potentially insulting or threatening topic;
- betrayal of confidences;
- losing the return envelope or return address;
- the wrong person may open the mail, and fail to pass it on to the most appropriate person.

Non-response can lead to responses that are systematically different (i.e. biased) than those from the whole sample or population, as the responses from those who did not respond might be distinctively different from those who actually responded.

Later in this chapter we discuss ways of improving response rates. However, here we wish to insert a note of caution: some researchers suggest that, for non-responders to an item, an average score for that item can be inserted. This might be acceptable if it can be shown that the sample or the population is fairly homogeneous, but, for heterogeneous populations or samples, or those where the variation in the sample or population is not known, it may be dangerous to assume homogeneity and hence to infer what the missing replies might have been, as this could distort the results.

Let us suppose that, out of a sample of 200 participants, 90 per cent reply (180 participants) to a 'yes/no' type of question, for example, for the question 'Do you agree with public examinations at age 11?', and let us say that 50 per cent (90 people) indicate 'yes' and 50 per cent indicate 'no'. If the 10 per cent who did not reply (20 people) had said 'yes' then this would clearly swing the results as 110 people say 'yes' (55 per cent) and 90 people say 'no' (45 per cent). However, if the response rates vary, then the maximum variation could be very different, as in Table 17.1 (cf. Fowler, 2009, p. 55). Table 17.1 assumes that, if 100 per cent had replied, 50 per cent said 'yes' and 50 per cent said 'no'; the rest of the table indicates the possible variation depending on response rate.

Table 17.1 indicates the possible variation in a simple 'yes/no' type of question. If a rating scale is chosen, for example a five-point rating scale, the number of options increases from two to five, and, correspondingly, the possibility for variation increases even further.

Improving response rates in a survey

A major difficult in survey research is securing a sufficiently high response rate to give credibility and reliability to the data. In some surveys, response rates can be as low as 20–30 per cent, and this compromises the reliability of the data very considerably. There is a difference between the *intended* and the *achieved* sample (Fogelman, 2002, p. 105). Punch (2003, p. 43) suggests that it is important to plan for poor response rates (e.g. by increasing the sample size) rather than trying to adjust sampling *post hoc*. He also suggests that access to the sample needs to be researched before the survey

TABLE 17.1 MAXIMUM VARIATION FOR LOW RESPONSE RATES IN A YES/NO QUESTION FOR A 50/50 DISTRIBUTION

Response rate (%)	*Variation in the true value of 'yes' and 'no' votes (lowest % to highest % in each category)*
100	50–50
90	45–55
80	40–60
70	35–65
60	30–70
50	25–75
40	20–80
30	15–85
20	10–90
10	5–95

commences, maybe pre-notifying potential participants if deemed desirable. He argues that a poor response level may also be due to the careless omission of details of how and when the questionnaire will be returned or collected. This is a matter that needs to be made clear in the questionnaire itself. In the case of a postal survey a stamped addressed envelope should always be included.

Kenett (2006) and Fowler (2009, p. 52) report that responses rates increase when people are interested in the subject matter of the survey, or if the subject is very relevant to them, or if completing the survey brings them a sense of satisfaction. Denscombe (2009b, p. 288) reports that response rates increase if the 'respondent burden' (the effort required by the respondent to answer a question) is low.

Further, the design, layout and presentation of the survey may also exert an influence on response rate. It is important to include a brief covering letter that explains the research clearly and introduces the researcher. The timing of the survey is important, for example, schools will not welcome researchers or surveys in examination periods or at special periods, for example, Christmas or inspection times (Fogelman, 2002, p. 106).

Finally, it is important to plan the follow-up to surveys, to ensure that non-respondents are called again and reminded of the request to complete the survey. Fowler (2009, p. 57) indicates that between a quarter and a third of people may agree to completing a survey if a follow-up is undertaken.

There are several possible ways of increasing response rates to mailed surveys (Aldridge and Levine, 2001; Diaz de Rada, 2005; Fowler, 2009, p. 56; Denscombe, 2014; Dillman *et al*., 2014; Williams *et al*., 2016), including:

- use follow-ups and polite reminders (e.g. by mail, email, telephone call) in which the reminder is short, polite, indicating the value of the respondent's participation and, if the reminder is postal, another clean copy of the questionnaire;
- use multiple and mixed modes of responding (i.e. avoid relying on a single mode, such as post, email, website, cellphone app, interview);
- give advance notification of the survey (e.g. by telephone, post or email);
- indicate how the survey is important and the benefits from it, and how (and what) the respondents can help in answering the survey;
- indicate the institutional affiliation (with a logo) that is sponsoring or supporting the survey and support for the survey from high-status or influential persons;
- provide information about the research through a covering letter and/or advance notification;
- avoid making the survey look like junk mail;
- thank the participants in advance;
- indicate that others have already answered the survey (do not be dishonest);
- give pre-paid stamped addressed envelopes for return of the survey;
- offer incentives for return (though increasing the financial incentive to a high figure does not bring commensurate returns in response rates);
- for a follow-up reminder, include a cover page, as this increases response rates;
- make it easy to answer the survey, keeping the respondent effort and burden to a minimum;
- make the questionnaire topic interesting, the design attractive and the questions interesting, clear and easy to answer, with easy-to-follow instructions and spacing of the text. Make instructions about responses and return very clear and easy;
- keep the survey short, easy to read and complete, and very clear;
- make response modes easy: giving too many kinds can lower response rates;
- avoid open-ended questions unless these are really important (as the quality of responses is usually poor to open-ended questions: people tend not to write anything or to write very little). Avoid placing open-ended questions at the start of a questionnaire;
- consider asking the respondents for an interview to complete the survey questionnaire;
- deliver the questionnaire personally rather than through mail;
- ensure that the questions or items are non-judgemental (e.g.in sensitive matters);
- avoid asking for sensitive or personal information unless it is absolutely necessary, particularly if asking for identifying features of children;
- indicate you own contact details, relevant and authentic professional information about yourself and how you can be reached;
- assure confidentiality, anonymity, privacy and security of information;
- send an email reminder to participants very shortly after the distribution of the survey.

Cooper and Schindler (2001, pp. 314–15) and Fowler (2009, p. 58) report that the following factors make little or no appreciable difference to response rates:

- personalizing the introductory letter;
- writing an introductory letter;
- promises of anonymity;

- questionnaire length (it is not always the case that a short questionnaire produces more returns than a long questionnaire, but researchers will need to consider the effect of a long survey questionnaire on the respondents – they may feel positive or negative about it, or set it aside temporarily and forget to return it later);
- size, reproduction and colour of the questionnaire;
- deadline dates for return (it was found that these did not increase response rate but did accelerate the return of questionnaires).

Potential respondents may be persuaded to participate depending on, for example:

- the status and prestige of the institution or researcher carrying out the research;
- the perceived benefit of the research;
- the perceived importance of the topic;
- personal interest in the research;
- interest in being interviewed, i.e. the interview experience;
- personal liking for, or empathy with, the researcher;
- feelings of duty to the public and sense of civic responsibility;
- loneliness or boredom (nothing else to do);
- sense of self-importance.

Dillman (2007) suggests that response rates can be increased if, in sequence: (a) non-respondents are sent a friendly reminder after ten days, stressing the importance of the research; (b) non-respondents are sent a further friendly reminder ten days after the initial reminder, stressing the importance of the research; (c) a telephone call is made to the respondents (if the number is known) shortly after the second reminder, indicating the importance of the research.

Fowler (2009, p. 60) suggests that the initial questionnaire might also include a statement to say that completion and return of the questionnaire will ensure that no follow-up reminders will be sent (though this may be regarded by some respondents as presumptuous).

17.8 Survey sampling

Sampling is a key feature of a survey approach, and we advise readers to look closely at Chapter 12 (sampling). Researchers must take sampling decisions early in the overall planning of a survey (see Figure 17.1) in light of the population from which they want to sample, and this involves, for example:

- identifying the target population (who, how large and what are their characteristics of interest?);
- deciding whether a sample or the whole population is necessary (e.g. it may be possible to have a whole population if access and size render it feasible, such as all the staff of a school);
- the sampling frame (all those to be included in the sample);
- the sampling strategy (probability and non-probability) and type of sample;
- sampling error;
- weighted samples for small groups (e.g. before the survey is conducted and post-stratification: after the survey has been conducted).

Often the researcher will not know the population size or heterogeneity of the characteristics of the population, and, in this event, it is advisable to have as large a sample as possible (see Chapter 12 for determining sample size).

We have already seen that due to factors of expense, time and accessibility, it is not always possible or practical to obtain measures from a population. Indeed Wilson *et al*. (2006, p. 352) draw attention to the tension between the need for large samples in order to conduct 'robust statistical analysis', and issues of resources such as cost and practicability (p. 353). Researchers endeavour, therefore, to collect information from a smaller group or subset of the population in such a way that the knowledge gained is representative of the total population under study, i.e. a sample. Unless researchers identify the total population in advance, it is virtually impossible for them to assess how representative the sample is which they have drawn. Chapter 12 addresses probability and non-probability samples, and we refer readers to the detailed discussion of these in that chapter. The researcher will need to decide the sampling strategy to be used on the basis of fitness for purpose, for example:

- a probability and non-probability sample;
- the desire to generalize, and to whom;
- the sampling frame (those who are eligible to be included);
- the sample size;
- the representativeness of the sample;
- access to the sample;
- the anticipated response rate.

Even if the researcher has taken extraordinary care with the sampling strategy, there may still be problems (e.g. response rate, respondent characteristics or availability) that can interfere with the best-laid strategies.

In addition to the sampling strategy to be used, there are the issues of sample size and selection. We discussed

this in Chapter 12, but here we wish to address the issue of practicability. For example, let us say that, in the interests of precision, the researcher wishes to have a sample in which there are four strata (e.g. age groups in a primary school), and that each stratum comprised 50 students, i.e. 200 students in total. If that researcher wished to increase the sample size of one stratum by, say, 20 students, this would necessitate an overall increase of 80 students (20×4) in the sample. Do the benefits outweigh the costs here?

An alternative to increasing the *total* size of the sample would be to increase the size of one stratum only, under certain conditions. For example, let us say that the researcher is studying attitudes of males and females to learning science, in a secondary school which had only recently moved from being a single-sex boys' school to a mixed sex school, so the ratio of male to female students is 4:1. The researcher wishes to include a minimum of 200 female students. This could require a total of 1,000 students in the sample (200 females + $\{200 \times 4\}$ male students in the sample); this could be unmanageable. Rather, the researcher could identify two female students for each male student (i.e. 400 females) and then, when analysing the data, could give one quarter of the weight to the response of the female students, in order to gain a truer representation of the target population of the school. This would bring the total sample to 600 students, rather than 1,000, involved in the survey. Oversampling a smaller group (in this case the females) and then weighting the analysis is frequently undertaken in surveys (cf. Fowler, 2009, p. 27).

In sampling, the probability might also exist of excluding some legitimate members of population in the target sample; however, the researcher will need to weigh the cost of excluding these members (e.g. the very hard to reach) against the cost of ensuring that they are included – the benefit gained from including them may not justify the time, cost and effort (cf. Fowler, 2009, p. 179). Similarly, the precision gained from stratified sampling (see Chapter 12) may not be worth the price to be paid in necessarily increasing the sample size in order to represent each stratum.

In many cases a sampling strategy may be in more than one stage. For example, let us consider the instance of a survey of 1,000 biology students from a population of 10,000 biology students in a city. In the first stage, a cluster group of, say, ten schools is identified (A), then, within that, a cluster by age group of students (B), and then, within that, the cluster of individuals in that group who are studying biology (C), and, finally, the sample (D) is taken from that group. The intention is to arrive at (D), but in order to reach this point a series of other steps has to be taken.

This raises the matter of deciding the steps to be taken. For example, the researcher could decide the sampling for the survey of the biology students by taking the random sample of 1,000 students from ten schools. The researcher lists all the 1,000 relevant students from the list of 10,000 students, and decides to select 100 students from each of the ten schools (a biology student, therefore, in one of these ten schools has a one in ten chance of being selected). Alternatively, the researcher could decide to sample from five schools only, with 200 students from each of the five schools, so students in each of these five schools have a one in five chance of being selected. Alternatively, the researcher could decide to sample from two schools, with 500 students, so students in each of these two schools have a one in two chance of being selected. There are other permutations. The point here is that, as the number of schools decreases, so does the possible cost of conducting the survey, but so does the overall reliability, as so few schools are included. It is a trade-off.

In order to reduce sampling error (the variation of the mean scores of the sample from the mean score of the population), a general rule is to increase the sample size, and this is good advice. However, it has to be tempered by the fact that the effect of increasing the sample size in a small sample reduces sampling error more than in a large sample, for example, increasing the sample size from 50 to 80 (30 persons) will have greater impact on reducing sampling error than increasing the sample size from 500 to 530 (30 persons). Hence it may be of little benefit simply to increase sample sizes in already-large samples.

The researcher has to exercise his or her judgement in attending to sampling. For example, if it is already known that a population is homogeneous, then the researcher may feel it a needless exercise in having too large and unmanageable a sample if the results are not likely to be much different from those of a small sample of the same homogeneous group (though theoretical sampling (see Chapter 37) may suggest where a researcher needs to include participants from other small samples). As Fowler (2009, p. 44) remarks, the results of a sample of 150 people will describe a population of 15,000 or 25 million with more or less the same degree of accuracy. He remarks that samples of more than 150 or 200 may yield only modest gains to the precision of the data (p. 45), though this, of course, has to be addressed in relation to the population characteristics, the number, size and kind of strata to be included, and the type of sample being used. Sampling

errors, he notes (p. 45) are more a function of sample size than of the proportions of the sample to the population. Further, he advocates probability rather than non-probability samples, unless there are persuasive reasons for non-probability samples to be used.

Whilst sample sizes can be calculated on the basis of statistics alone (e.g. confidence levels, confidence intervals, population size, statistical power and so on, see Chapter 12), this is often not the sole criterion, as it accords a degree of precision to the sample which takes insufficient account of other sampling issues, for example, access, variation or homogeneity in the population, levels of literacy in the population (e.g. in the case of a self-administered questionnaire survey), number and type of variables and costs.

Sampling is one of several sources of error in surveys, as indicated earlier in this chapter.

17.9 Longitudinal and cross-sectional surveys

The term 'longitudinal' describes a variety of studies that are conducted over a period of time. A clear distinction is drawn between longitudinal and cross-sectional studies. The longitudinal study gathers data over an extended period of time: a short-term investigation may take several weeks or months; a long-term study can extend over many years. Where successive measures are taken at different points in time from the same respondents, the term 'follow-up study' or 'cohort study' is used in the British literature, the equivalent term in the US being the 'panel study'. The term 'cohort' is a group of people with some common characteristic.

Where different respondents are studied at one or more different points in time, the study is called 'cross-sectional', i.e. a cross-section of the population is taken to investigate the topic(s) of interest. Where a few selected factors are studied continuously over time, the term 'trend study' is employed. One example of regular or repeated cross-sectional social surveys is the General Household Survey, in which the same questions are asked every year, though they are put to a different sample of the population each time. The British Social Attitudes Survey is an example of a repeated cross-sectional survey, using some 3,600 respondents.

A famous example of a longitudinal (cohort) study is the UK's National Child Development Study, which started in 1958. The British General Household Panel Survey interviewed individuals from a representative sample each year in the 1990s. Another example is the British Family Expenditure Survey. These latter two are cross-sectional in that they tell us about the population at a given point in time, and hence provide aggregated data.

By contrast, longitudinal studies can also provide individual-level data, by focusing on the same individuals over time (e.g. the Household Panel Studies which follow individuals and families over time (Ruspini, 2002, p. 4). Lazarsfeld introduced the concept of a panel in the 1940s, attempting to identify causal patterns and the difficulties in tracing these (Ruspini, 2002, p. 13)).

Longitudinal studies

Longitudinal studies can be of the survey type or of other types (e.g. case study). Here we confine ourselves to the survey type. Longitudinal studies can include trend studies, cohort studies and panel studies (Creswell, 2012), and we discuss these below. A useful centre for longitudinal studies in education is at the University of London: www.cls.ioe.ac.uk/default.aspx.

Longitudinal studies can use repeated cross-sectional studies, which are conducted regularly, each time with a largely different sample or, indeed, an entirely new sample (Ruspini, 2002, p. 3), or use the same sample over time. They enable researchers to: 'analyse the duration of social phenomena' (p. 24); highlight similarities, differences and changes over time in respect of one or more variables or participants (within and between participants); identify long-term ('sleeper') effects; and explain changes in terms of stable characteristics, for example sex, or variable characteristics, such as income. The appeal of longitudinal research is its ability to establish causality and to make inferences. Ruspini adds to these the ability of longitudinal research to 'construct more complicated behavioural models than purely cross-sectional or time-series data' (p. 26); they can catch the complexity of human behaviour. Further, longitudinal studies can combine numerical and qualitative data.

Retrospective analysis is not confined to longitudinal studies alone. For example, Rose and Sullivan (1993, p. 185) and Ruane (2005, p. 87) suggest that cross-sectional studies can use retrospective factual questions, for example, previous occupations, dates of birth within the family, dates of marriage and/or divorce, though Rose and Sullivan (1993, p. 185) advise against collecting other types of retrospective data in cross-sectional studies, as the quality (e.g. reliability) of the data diminishes the further back one asks respondents to recall previous states or even facts.

It is important in longitudinal studies to decide when and how frequently to collect data over time, and this is informed by issues of fitness for purpose as well as practicability. Further, in order to allow for attrition

(dropout) of the sample, it is wise to have as large a sample as practicable and possible at the start of the study (Wilson *et al*., 2006, p. 354).

Cohort studies

A cohort study focuses on a specific population in which all its members have the specific defining characteristic that is of interest to the researcher (e.g. the National Child Development Study in the UK; the Millennium Cohort Study). In a cohort study the specific population is tracked over a specific period of time but selective sampling within that sample occurs. This means that different members of a cohort are included each time. For example, the population might be eighteen-year-olds in the UK; at one time point (say, when they are twenty-five years old) the population might be sampled, and then at another time point (say, when they are thirty-five) the same population might be sampled but different members of the population will be in the sample.

Cohort studies and trend studies can be *prospective* longitudinal methods, in that they are ongoing in their collection of information about individuals or their monitoring of specific events. *Retrospective* longitudinal studies, on the other hand, focus upon individuals who have reached some defined end-point or state. For example, a group of young people may be the researcher's particular interest (intending social workers, convicted drug offenders or university dropouts, for example), with questions such as: 'Is there anything about your previous experience that can account for your present situation?' Retrospective longitudinal studies will specify the period over which to be retrospective, for example, one year, five years.

Panel studies

In contrast to a cohort study, in a panel study exactly the same individuals are tracked over time. An example of this is the Panel Study of Income Dynamics in the US. Another example from the UK is the '7 Up' study which started in 1964 and tracks a small group of individuals every seven years, yielding insights into social and cultural stratification, reproduction and the self-fulfilling prophecy.

Whilst this type of study has the attraction of tracking the same people over time, this same requirement also has its disadvantages in terms of keeping contact with those individuals and addressing attrition. Panel studies are useful for investigating causality and change over time.

Trend studies

Trend studies focus on factors (e.g. mathematics performance) rather than people, and these factors are studied over time. New samples – different people – are drawn at each stage of the data collection, but focus on the same factors, and if random samples are used, they can be representative of the wider population. By taking different samples the problem of reactivity is reduced (see below: 'pre-test sensitisation'), i.e. earlier surveys affecting the behaviour of participants in the later surveys. This is particularly useful if the research is being conducted on sensitive issues, as raising a sensitive issue early on in the research may change an individual's behaviour, which could affect the responses in a later round of data collection. By drawing a different sample each time, this problem is overcome.

Trend or prediction studies have an obvious importance to educational administrators or planners. Like cohort studies, they can be of relatively short or long duration. Essentially, the trend study examines recorded data to establish patterns of change that have already occurred in order to predict what will be likely to occur in the future. In trend studies, two or more cross-sectional studies are undertaken with identical age groups at more than one point in time in order to make comparisons over time (e.g. the Scholastic Aptitude and Achievement tests in the US and the National Assessment of Educational Progress results). A major difficulty that researchers face in conducting trend analyses is the intrusion of unpredictable factors that invalidate forecasts formulated on past data. For this reason, short-term trend studies tend to be more accurate than long-term analyses. Trend studies do not include the same respondents over time, so the possibility exists for variation in data due to the different respondents rather than the change in trends. Gorard (2001b, p. 87) suggests that this problem can be attenuated by a 'rolling sample' in which a proportion of the original sample is retained in the second wave of data collection, and a proportion of this sample is retained in the third wave, and so on.

Cross-sectional studies

A cross-sectional study is one that produces a 'snapshot' of a population at one particular point in time. The epitome of the cross-sectional study is a national survey in which a representative sample of the population consisting of individuals of different ages, different occupations, different educational and income levels, and residing in different parts of the country, is interviewed on the same day. In education, cross-sectional studies can involve indirect measures of the nature and rate of changes in the physical and intellectual development of samples of children drawn from representative age levels. The single 'snapshot' of the cross-sectional study provides researchers with data for either a retrospective or a prospective enquiry.

A cross-sectional study can also bear several hallmarks of a longitudinal study of parallel groups (e.g. age groups) which are drawn simultaneously from the population. For example, drawing students aged five, seven, nine and eleven at a single point in time would bear some characteristics of a longitudinal study in that developments over age groups could be seen, though, of course, it would not have the same weight as a longitudinal study conducted on the same age group over time. This is the case for international studies of educational achievement, requiring samples to be drawn from the same population (Lietz and Keeves, 1997, p. 122) and for factors that might influence changes in the dependent variables to remain constant across the age groups.

Cross-sectional studies, catching a frozen moment in time, may be ineffective for studying change or causality. If changes are to be addressed through cross-sectional surveys, then this suggests the need for repeated applications of the survey, or the use of trend analysis.

The main types of longitudinal study are illustrated in Figure 17.2.

17.10 Strengths and weaknesses of longitudinal, cohort and cross-sectional studies

Longitudinal studies of the cohort analysis type have an important place in the armoury of educational researchers. Longitudinal studies have considerable potential for yielding rich data that can trace changes over time, and with great accuracy (Gorard, 2001b, p. 86). On the other hand, they suffer from problems of attrition (participants leaving the research over time, a particular problem in panel studies which research the same individuals over time), and they can be expensive to conduct in terms of time and money (Ruspini, 2002, p. 71). Gorard (2001b) reports a study of careers and identities that had an initial response rate of between 60 and 70 per cent in the first round, and then risked dropping to 25 per cent by the third round, becoming increasingly more middle class in each wave of the study; the same publication discusses a Youth Cohort Study in which only 45 per cent of the respondents took part in all three waves of the data collection. Ruspini (2002, p. 72) identifies an attrition rate of 78 per cent in the three waves of the European Community Household Panel survey of the UK in 1997.

Ruspini also indicates how a small measurement error in a longitudinal study may be compounded over time. She gives the example of an error in income occurring at a point in time (p. 72) that could lead to 'false transitions' appearing over time in regard to poverty and unemployment.

Further, long-term studies, Gorard (2001b, p. 86) avers, face 'a threat to internal validity' that stems from the need 'to test and re-test the same individuals'. Dooley (2001, p. 120) terms this 'pre-test sensitisation'; it is also termed 'panel conditioning' or 'time-in sample bias' (Ruspini, 2002, p. 73). Here the first interview in an interview survey can cause changes in the second interview, i.e. the first interview might set up a self-fulfilling prophecy that is recorded in the second interview. He gives the example of a health survey in the first round of data collection, which may raise

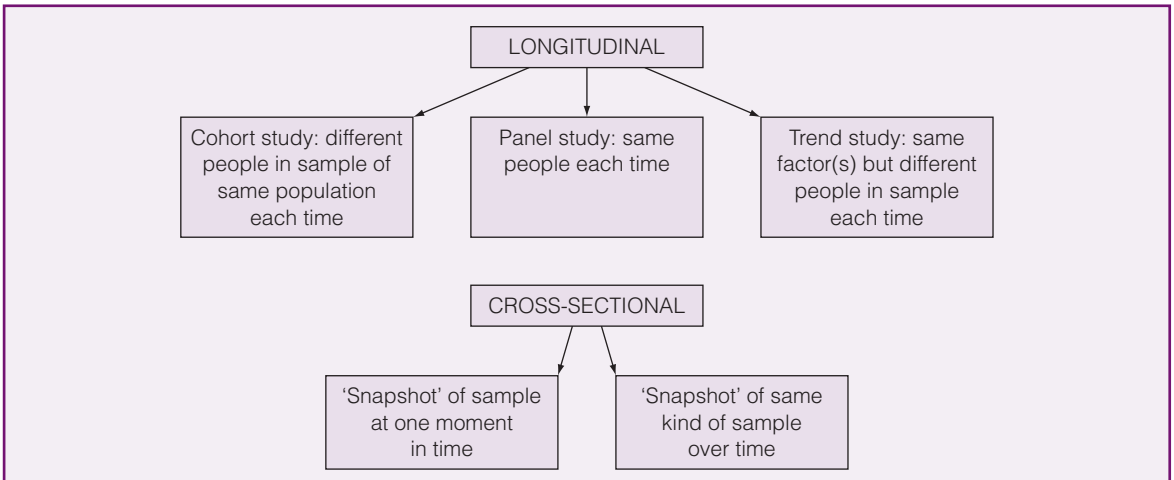


FIGURE 17.2 Types of survey

participants' awareness of the dangers of smoking, such that they reduce or give up smoking by the time the second round takes place. Trend studies overcome this problem by drawing different populations at each round of data collection.

Dooley (2001) also identifies difficulties caused by changes in the research staff over time in longitudinal surveys. Changes in interviewee response, he suggests, may be due to having different researchers rather than to the respondents themselves. Even using the same instruments, different researchers may use them differently (e.g. in interviewing behaviour).

To add to these matters, Ruspini (2002, p. 73) suggests that longitudinal data are affected by:

- history (events occurring may change the observations of a group under study);
- maturation (participants mature at different speeds and in different ways);
- testing (test sensitization may occur – participants learn from exposure to repeated testing/interviews);
- the timing of cause and effect (some causes may produce virtually instantaneous effects and others may take a long time for the effects to show);
- the direction of causality not always being clear or singular.

A major concern in longitudinal studies concerns the comparability of data over time. For example, though public examinations may remain constant over time (e.g. GCSE, A levels), the contents and format of those examinations do not. (This rehearses the argument that public examinations are becoming easier over time.) This issue concerns the need to ensure consistency in the data-collection instruments over time. Further, if comparability of data in a longitudinal study is to be addressed then this means that the initial rounds of data collection will need to anticipate and include all the variables that will be addressed over time.

Longitudinal studies are more prone to attrition than cross-sectional studies, and are more expensive to conduct in terms of time and cost. On the other hand, whereas trend studies change their populations, thereby disabling micro-level – individual-level – analysis from being conducted, longitudinal analysis enables such individual-level analysis to be performed. Indeed whereas cross-sectional designs (even if they are repeated cross-sectional designs) may be unsuitable for studying developmental patterns and causality within cohorts, in longitudinal analysis this is a strength. Longitudinal data can supply 'satisfactory answers to questions concerning the dynamics and the determinants of individual behaviour' (Ruspini, 2002, p. 71), issues which are not easily addressed in cross-sectional designs.

Retrospective longitudinal studies rely on participants' memories which may be faulty; the further back one's memory reaches, the greater is the danger of distortion or inability to recall. Memory is affected by, for example (Ruspini, 2002, p. 97):

- the time that has elapsed since the event took place;
- the significance of the event for the participant;
- the amount of information required for the study – the greater the amount, the harder it is to provide;
- the contamination/interference effect of other memories of a similar event (i.e. the inability to separate similar events);
- the emotional content or the social desirability of the content;
- the psychological condition of the participant at interview.

Further, participants will look at past events through the lens of hindsight and subsequent events rather than what those events meant at the time. Moreover, it is not always easy for these participants to recall their emotional state at the time in question. Factually speaking, it may not be possible to gather data from some time past, as they simply do not exist (e.g. medical records, data on income) or they cannot be found, recovered or accessed.

Cohort studies of human development conducted on representative samples of populations are uniquely able to identify typical patterns of development and to reveal factors operating on those samples which elude other research designs. They permit researchers to examine individual variations in characteristics or traits, and to produce individual development curves. Cohort studies, too, are particularly appropriate when investigators attempt to establish causal relationships, as this involves identifying changes in certain characteristics which result in changes in others.

Cross-sectional designs are inappropriate in causal research as they cannot sustain causal analysis unless they are repeated over time, as causality has a necessary time dimension. Cohort analysis is especially useful in sociological research because it can show how changing properties of individuals fit together into changing properties of social systems as a whole. For example, the study of staff morale and its association with the emerging organizational climate of a newly opened school would lend itself to this type of developmental research. A further strength of cohort studies in schools is that they provide longitudinal records whose

value takes account of the known fallibility of any single test or assessment. Finally, time, often a limiting factor in experimental and interview settings, is generally more readily available in cohort studies, allowing the researcher greater opportunity to observe trends and to distinguish 'real' changes from chance occurrences (see Bailey, 1994).

In longitudinal, cohort and trend studies the characteristics of respondents are likely to affect results (Robson, 1993, p. 128). For example, their memory, knowledge, motivation and personality may affect their responses, and indeed they may withhold information, particularly if it is sensitive.

Longitudinal research indicates the influence of biological factors over time (e.g. human development), environmental influences and intervention influences (Keeves, 1997a, p. 139) and their interactions. Addressing these, the appeal of longitudinal analysis is that it enables researches to conduct causal analysis. Time series studies in longitudinal research also enable emergent patterns to be observed over time, by examining a given range of variables over time, in addition to other factors. This enables individual and group profiles to be examined over time and development, indicating similarities and differences within and between individuals and groups in respect of given variables.

Longitudinal studies suffer several disadvantages (though the gravity of these weaknesses is challenged by supporters of cohort analysis). The disadvantages are, first, that they are time-consuming and expensive, because the researcher is obliged to wait for growth data to accumulate. Second, there is the difficulty of sample mortality. Inevitably during the course of a long-term cohort study, subjects drop out, are lost or refuse further cooperation. Such attrition makes it unlikely that those who remain in the study are as representative of the population as the original sample. Sometimes attempts are made to lessen the effects of sample mortality by introducing aspects of cross-sectional study design, that is, 'topping up' the original cohort sample size at each time of retesting with the same number of respondents drawn from the same population. The problem here is that differences arising in the data from one survey to the next may then be accounted for by differences in the persons surveyed rather than by genuine changes or trends.

A third difficulty has been termed the 'control effect' (sometimes referred to as 'measurement effect'). Often, repeated interviewing results in an undesired and confusing effect on the actions or attitudes under study, influencing the behaviour of subjects, sensitizing them to matters that have hitherto passed unnoticed, or stimulating them to communicate with others on unwanted topics (see Riley, 1963). Fourth, cohort studies can suffer from the interaction of biological, environmental and intervention influences (Keeves, 1997a, p. 139). Finally, cohort studies in education pose considerable problems of organization due to the continuous changes that occur in pupils, staff, teaching methods and the like. Such changes make it highly unlikely that a study will be completed in the way that it was originally planned.

Cohort studies, as we have seen, are particularly appropriate in research on human growth and development. Why then are so many studies cross-sectional rather than cohort studies? The reason is that they have a number of advantages over cohort studies: they are less expensive; they produce findings more quickly; they are less likely to suffer from control effects; and they are more likely to secure the cooperation of respondents on a 'one-off' basis. Generally, cross-sectional designs are able to include more subjects than are cohort designs.

The strengths of cohort analysis are the weaknesses of the cross-sectional design. The cross-sectional study is a less effective method for the researcher who is concerned to identify individual variations in growth or to establish causal relationships between variables. Sampling in a cross-sectional study is complicated because different subjects are involved at each age level and may not be comparable. Further problems arising out of selection effects and obscuring irregularities in growth weakens the cross-sectional study so much that one observer dismisses the method as a highly unsatisfactory way of obtaining developmental data except for the crudest purposes.

Douglas (1976a), who pioneered the first national cohort study in any country, makes a spirited defence of the method against the common criticisms that are levelled against it – that it is expensive and time-consuming. His account of the advantages of cohort analysis over cross-sectional designs is summarized in Box 17.1.

Cross-sectional studies require attention to sampling in order to ensure that the information on which the sample is based is comprehensive (Lietz and Keeves, 1997, p. 124). Further, there is a risk that some potential participants may decline to take part, thereby weakening the sample, or some may not answer specific questions or, wittingly or unwittingly, give incorrect answers. Measurement error may also occur if the instrument is faulty, for example, using inappropriate metrics or scales.

The comparative strengths and weaknesses of longitudinal studies (including retrospective studies),

BOX 17.1 ADVANTAGES OF COHORT OVER CROSS-SECTIONAL DESIGNS

1. Some types of information, for example, on attitudes or assessment of potential ability, are only meaningful if collected contemporaneously. Other types are more complete or more accurate if collected during the course of a longitudinal survey, though they are likely to have some value even if collected retrospectively, for example, length of schooling, job history, geographical movement.
2. In cohort studies, no duplication of information occurs, whereas in cross-sectional studies the same type of background information has to be collected on each occasion. This increases the interviewing costs.
3. The omission of even a single variable, later found to be important, from a cross-sectional study is a disaster, whereas it is usually possible in a cohort study to fill the gap, even if only partially, in a subsequent interview.
4. A cohort study allows the accumulation of a much larger number of variables, extending over a much wider area of knowledge than would be possible in a cross-sectional study. This is of course because the collection can be spread over many interviews. Moreover, information may be obtained at the most appropriate time, for example, information on job entry may be obtained when it occurs even if this varies from one member of the sample to another.
5. Starting with a birth cohort removes later problems of sampling and allows the extensive use of subsamples. It also eases problems of estimating bias and reliability.
6. Longitudinal studies are free of one of the major obstacles to causal analysis, namely, the reinterpretation of remembered information so that it conforms to conventional views on causation. It also provides the means to assess the direction of effect.

Source: Adapted from Douglas (1976b)

cross-sectional analysis and trend studies are summarized in Table 17.2 (see also Rose and Sullivan, 1993, pp. 184–8).

Several of the strengths and weaknesses of retrospective longitudinal studies share the same characteristics as those of *ex post facto* research, discussed in Chapter 20.

17.11 Postal, interview and telephone surveys

Postal surveys

There are strengths and difficulties with postal and interview surveys. Postal surveys can reach a large number of people, gather data at comparatively low cost and quite quickly, and can give assurances of confidentiality (Robson, 1993; Bailey, 1994, p. 148; Dillman *et al*., 2014). Similarly they can be completed at the respondents' own convenience and in their preferred surroundings and own time; this can enable them to check information, if necessary (e.g. personal documents), and think about responses. As standardized wording is used, there is a useful degree of comparability across the responses, and, as no interviewer is present, there is no risk of interviewer bias. Further, postal questionnaires enable widely scattered populations to be reached.

Postal surveys can also be used to gather detailed sensitive qualitative data (Beckett and Clegg, 2007), not least because the non-presence of another person (e.g. an interviewer) can increase the honesty and richness of the data, whereas the presence of an interviewer might inhibit the respondent. Further, in a postal survey, the relations of power between the researcher and the respondent are often more equal than in an interview situation (in which the former often controls the situation more than the latter) (p. 308).

On the other hand, postal surveys typically suffer from a poor response rate, even though Dillman *et al*. (2014) comment they have moved from having the lowest response rate to having response rates higher than telephone surveys. Mailed surveys are reported to have an approximately 20 per cent response rate, which is far lower than telephone and face-to-face surveys (Colorado State University, 2016). Diaz de Rada and Dominguez (2015) note that postal surveys feature greater acquiescence than other kinds of survey, with more unanswered questions.

Because researchers may not have any information about non-respondents, they may not know whether the sample is representative of the wider population. Further, respondents may not take the care required to complete the survey carefully, and, indeed, may misunderstand the questions. There is no way of checking this. Bailey (1994, p. 149) suggests that the very issues that make postal surveys attractive might also render them less appealing, for example:

TABLE 17.2 THE CHARACTERISTICS, STRENGTHS AND WEAKNESSES OF LONGITUDINAL, CROSS-SECTIONAL, TREND ANALYSIS AND RETROSPECTIVE LONGITUDINAL STUDIES

Study type	*Features*	*Strengths*	*Weaknesses*
Longitudinal studies (cohort/panel studies)	1 Single sample over extended period of time. 2 Enables the same individuals to be compared over time (diachronic analysis). 3 Micro-level analysis.	1 Useful for establishing causal relationships and for making reliable inferences. 2 Shows how changing properties of individuals fit into systemic change. 3 Operates within the known limits of instrumentation employed. 4 Separates real trends from chance occurrence. 5 Brings the benefits of extended time frames. 6 Useful for charting growth and development. 7 Gathers data contemporaneously rather than retrospectively, thereby avoiding the problems of selective or false memory. 8 Economical in that a picture of the sample is built up over time. 9 In-depth and comprehensive coverage of a wide range of variables, both initial and emergent – individual specific effects and population heterogeneity. 10 Enables change to be analysed at the *individual/micro*-level. 11 Enables the dynamics of change to be caught, the flows into and out of particular states and the transitions between states. 12 Individual level data are more accurate than macro-level, cross-sectional data. 13 Sampling error reduced as the study remains with the same sample over time. 14 Enables clear recommendations for intervention to be made.	1 Time-consuming – it takes a long time for the studies to be conducted and the results to emerge. 2 Problems of sample mortality heighten over time and diminish initial representativeness. 3 Control effects – repeated interviewing of the same sample influences their behaviour. 4 Intervening effects attenuate the initial research plan. 5 Problem of securing participation as it involves repeated contact. 6 Data, being rich at an individual level, are typically complex to analyse.

continued

TABLE 17.2 CONTINUED

Study type	*Features*	*Strengths*	*Weaknesses*
Cross-sectional studies	1 Snapshot of different samples at one or more points in time (synchronic analysis). 2 Large-scale and representative sampling. 3 Macro-level analysis. 4 Enables different groups to be compared. 5 Can be retrospective and/or prospective.	1 Comparatively quick to conduct. 2 Comparatively cheap to administer. 3 Limited control effects as subjects only participate once. 4 Stronger likelihood of participation as it is for a single time. 5 Charts aggregated patterns. 6 Useful for charting population-wide features at one or more single points in time. 7 Enable researchers to identify the proportions of people in particular groups or states. 8 Large samples enable inferential statistics to be used, e.g. to compare sub-groups within the sample.	1 Do not permit analysis of causal relationships. 2 Unable to chart individual variations in development or changes, and their significance. 3 Sampling not entirely comparable at each round of data collection as different samples are used. 4 Can be time-consuming as background details of each sample have to be collected each time. 5 Omission of a single variable can undermine the results significantly. 6 Unable to chart changing social processes over time. 7 They only permit analysis of overall, *net* change at the macro-level through aggregated data.
Trend analysis	1 Selected factors studied continuously over time. 2 Uses recorded data to predict future trends.	1 Maintains clarity of focus throughout the duration of the study. 2 Enables prediction and projection on the basis of identified and monitored variables and assumptions.	1 Neglects influence of unpredicted factors. 2 Past trends are not always a good predictor of future trends. 3 Formula-driven, i.e. could be too conservative or initial assumptions might be erroneous. 4 Neglects the implications of chaos and complexity theory, e.g. that long-range forecasting is dangerous. 5 The criteria for prediction may be imprecise.
Retrospective longitudinal studies	1 Retrospective analysis of history of a sample. 2 Individual- and micro-level data.	1 Useful for establishing causal relationships. 2 Clear focus (e.g. how did this particular end state or set of circumstances come to be?). 3 Enables data to be assembled that are not susceptible to experimental analysis.	1 Remembered information might be faulty, selective and inaccurate. 2 People might forget, suppress or fail to remember certain factors. 3 Individuals might interpret their own past behaviour in light of their subsequent events, i.e. the interpretations are not contemporaneous with the actual events. 4 The roots and causes of the end state may be multiple, diverse, complex, unidentified and unstraightforward to unravel. 5 Simple causality is unlikely. 6 A cause may be an effect and vice versa. 7 It is difficult to separate real from perceived or putative causes. 8 It is seldom easily falsifiable or confirmable.

- the standardization of wording;
- the inability to catch anything other than a verbal response;
- the lack of control over the environment in which the survey questionnaire is completed;
- the lack of control over the order in which the questions are read and answered;
- the risk that some questions will not be answered;
- the inability to record spontaneous answers;
- the difficulty in separating non-response from bad response, the former being where the intended respondent receives the survey but does not reply to it, and the latter being where the intended recipient does not receive the survey, for example, because she/he has moved house;
- the need for simplicity in format as there is no interviewer present to guide the respondent through a more complex format.

Postal surveys are an example of self-administered surveys. The anonymity and absence of face-to-face interaction between the interviewer and the respondent can render these useful for asking sensitive questions (Strange *et al*., 2003, p. 337), though Fowler (2009, p. 74) also counsels that sensitive questions can sometimes be handled better in private face-to-face interviews. In self-administered surveys, Fowler (2009, p. 72) remarks that it is advisable to keep to closed questions and make the response categories simple and explicit (e.g. ticking a box). If open questions are to be asked then, he indicates, it is better to gather the survey data in a face-to-face interview.

Further, Diaz de Rada (2005) reports that the design, size and colour of the paper used in postal surveys affects response rates. Small-sized questionnaires were mostly returned by males and those under sixty-four years of age (p. 69), whilst larger-sized questionnaires were mostly returned by females and those over the age of sixty-five (p. 70). He recommends using paper size 14.85 × 21 cm (i.e. a sheet of A4-sized paper folded in half), white paper, and including a cover page (p. 73) (though this inevitably increases the number of pages in a questionnaire, and this can be off-putting for respondents). He reports that paper size has no effect on the quality of the responses.

Interview surveys

Whereas postal surveys are self-administered, interview surveys are supervised and hence potentially prone to fewer difficulties. Interview methods of gathering survey data are useful in that the presence of the interviewer can help clarify queries from the respondents and can stimulate the respondent to give full answers to an on-the-spot researcher rather than an anonymous researcher known only through an introductory letter (Robson, 1993). Indeed face-to-face encounters can improve response rates. Further, as interviews can be flexible, questioners are able both to probe and to explain more fully (Bailey, 1994, p. 174). Interviews are also useful when respondents have problems with reading and writing. Using non-verbal behaviour to encourage respondents to participate is also possible. Moreover, with interviews there are greater opportunities to control the environment in which the survey is conducted, particularly in respect of privacy, noise and external distractions.

The effective interviewer, Fowler (2009, p. 128) claims, is business-like and assertive whilst being engaging, friendly and kind. Fowler argues for great care with choosing interviewers and training them, as much can hang on their behaviour.

The potential for trust, rapport and cooperation between the interviewer and the respondent is strong in face-to-face encounters (Dooley, 2001, p. 122; Gwartney, 2007, p. 16). Further, interviewers can either ensure that the sequence of the survey protocol is strictly adhered to or they can tailor the order of responses to individual participants, making certain that all questions are answered. Interview surveys, moreover, can guarantee that it is the respondent alone who answers the questions, whereas in postal surveys the researcher never knows what help or comments are solicited from, or given by, other parties. Bailey (1994) adds that the opportunity for spontaneous behaviour and responses is also possible in interview surveys, and interviews can use more complex structures than postal questionnaires, the researcher being on hand to take participants through the items.

On the other hand, the very features which make interview methods attractive may also make them problematic. For example, interview survey methods may be affected by the characteristics of the interviewer (e.g. sex, race, age, ethnicity, personality, skills, perceived social status, clothing and appearance). They may also be affected by the conduct of the interview itself (e.g. rapport between the interviewer and the interviewee), and interviewees may be reluctant to disclose some information if they feel that the interview will not be anonymous or if sensitive information is being requested. The flexibility which the interview gives also contributes to the potential lack of standardization of the interview survey, and this may render consistency, and thereby reliability, a problem.

Interview surveys are costly in time for the researcher and the interviewee, and, as they are conducted at a fixed time, they may prevent the interviewee

from consulting records that may be important to answer the questions. Further, they may require the interviewer to travel long distances to reach interviewees, which can be expensive both in time and travel costs (Bailey, 1994, p. 175). If interviews are intended to be conducted in the participants' own homes, then participants may be unwilling to admit strangers. Moreover, neighbourhoods may be dangerous for some researchers to visit (e.g. a white researcher with a clipboard going into a non-white area of great deprivation, or a black researcher going into a conservative white area).

Telephone surveys

Telephone surveys are located between mailed questionnaires and personal interviews (Arnon and Reichel, 2009). Dillman *et al*. (2014) note the rapid decline in telephone interviewing (p. 11) with the reduction in landlines, the rise in cellphones, the lack of listing of call numbers and the rise in screening callers. However, telephone interviews have the attraction of overcoming bias in the researcher or the interviewee that may be caused by social characteristics or matters of age, dress, race, ethnicity, appearance etc. (e.g. Gwartney, 2007, p. 16). Indeed Denscombe (2014) suggests that people are 'more honest and open' on the phone than in a postal questionnaire (p. 12).

Telephone surveys require the interviewer to be an articulate, clear speaker and a good listener, and able to key in interviewee responses onto a computer whilst listening and speaking (Denscombe, 2014, pp. 42–3). They have the advantage of reducing costs in time and travel, for when a potential respondent is not at home, a call-back is cheap and the time to redial is short (Dooley, 2001, p. 122; Arnon and Reichel, 2009, p. 179), and, using Internet services such as Skype, telephone surveys can be almost free of charge and include face-to-face viewing. Revisits to often distant locations, on the other hand, can incur considerable expense in time and travel. Furthermore, if the intended participant is unable or unwilling to respond, then it is a relatively easy matter to maintain the required sample size by calling a replacement. Again, where respondents are unable or unwilling to answer all the questions required, then their partial replies may be discarded and further substitutes sought from the sample listing. It is easy to see why telephone interviews must always have a much longer list of potential respondents in order to attain the required sample size.

Not everyone has a telephone (e.g. the poor, the young) and this may lead to a skewed sample (Arnon and Reichel, 2009, p. 179). Nor, for that matter, is everyone available for interview, particularly if they work. Further, many people are 'ex-directory', i.e. their numbers are withheld from public scrutiny. In addition, Dooley (2001, p. 123) reports that younger, single and higher occupational status groups use electronic facilities that screen out and delete researchers' calls and these could lead to a skewed sample. Indeed Fowler (2009, p. 75) indicates that telephone surveys tend to elicit more socially desirable answers than face-to-face interviews.

Even when the telephone is answered, the person responding may not be the most suitable one to take the call; she/he may not know the answer to the questions or have access to the kind of information required. For example, in an inquiry about household budgets, the respondent may simply be ignorant about a family's income or expenditure on particular items. A child may answer the call, or an elderly person who may not be the householder. Interviewers will need to prepare a set of preliminary screening questions or arrange a call-back time when a more appropriate person can be interviewed.

Telephone interviewing has its own strengths and weaknesses. For example, more often than not a respondent's sex will be clear from their voice, so some questions may be unnecessary or inappropriate. On the other hand, it is unwise to have several multiple choices in a telephone interview, as respondents will simply forget the categories available, there being no written prompts to which the respondent can refer.

Similarly, order effects can be high: items appearing early in the interview exert an influence on responses to later ones, whilst items appearing early in a list of responses may be given greater consideration than those occurring later, a matter not confined to telephone surveys but to questionnaires in general. Dooley (2001, p. 136) indicates a 17 per cent difference in agreement recorded to a general statement question when it appeared *before* rather than *after* a specific statement, and other research demonstrates that responses to particular questions are affected by questions surrounding them. His advice is to ask *general* questions before *specific* ones, otherwise the general questions are influenced by earlier responses to specific questions. Once again, this is a matter not confined to telephone surveys but to questionnaires in general.

Further, if the questioning becomes too sensitive, respondents may simply hang up in the middle of the survey interview, tell lies or withhold information. Dooley (2001, p. 123) reports that, in comparison to face-to-face interviews, telephone respondents tend to produce more missing data, to be more evasive, more acquiescent (i.e. they tend to agree more with statements) and more extreme in their responses (e.g. opting for the extreme ends of rating scales).

Fowler (2009, pp. 73–4) also indicates that, in a telephone survey, it is unwise to have too many response scale points, that it is better to avoid long lists of items and that it is advisable to read the statement before indicating the response categories, unless a long list of items is to be given (i.e. is unavoidable), in which case he suggests that it is better to read and re-read the response categories to the respondent before the list of statements. All of these points take account of the limits of the short-term memories on which respondents often rely in a telephone interview. He also suggests (p. 73) that complex questions can be approached in a staged manner. For example, if a researcher wishes to ask about a ten-category item (e.g. income level of the teacher), then the researcher could start with a general question (e.g. above or below a particular figure), and then, once that category has been identified, proceed to a sub-category, for example, between such-and-such a figure; this avoids overload of asking a respondent to remember ten categories.

Because telephone interviews lack the sensory stimulation of visual or face-to-face interviews or written instructions and presentation, it is unwise to plan a long telephone survey call. Ten to fifteen minutes is often the maximum time tolerable to most respondents, and indeed fifteen minutes for many people is too long. This means that careful piloting must take place in order to include those items, and only those items, that are necessary for the research. The risk to reliability and validity is considerable, as the number of items may be fewer than in other forms of data collection.

Procedures for telephone interviews also need to be decided (Gwartney, 2007), for example:

- how many times to let the telephone ring before conceding that there is nobody to answer the call (Gwartney (2007, p. 99) suggests eight rings);
- how to introduce the caller and the project;
- what to say and how to introduce items and conduct the interview;
- how to determine who is receiving the call and whether he/she is the appropriate person to answer the call;
- whether to leave a message on an answerphone/voicemail/call-back facility and, if so, what that message will be;
- how to handle language problems (e.g. which language is being used, meanings/explanations/vocabulary);
- how to handle the situation if the receiver asks to call back later;
- what to say and how to control the caller's voice/tone/pitch/speed/pace of questions/repetitions/language/intonation/register;
- the caller's pronunciation, enunciation and reading out loud;
- the caller's ability to clarify, summarize, reiterate, probe (and when to stop probing), prompt (if the receiver does not understand), confirm, affirm, respond, give feedback, encourage respondents, keep respondents focused and to the point;
- how to conduct closed and open questions, sensitive, factual and opinion-based questions;
- how to indicate the nature and format of the responses sought;
- the caller's ability to handle the called person's initial hostility, refusal, reluctance to take part, feelings of invasion of privacy, lack of interest, reluctance to disclose information, feelings of being harassed or singled out, anger, antagonism, lack of interest, incomplete answers, hurriedness to complete, slowness or hesitancy, mistrust, rudeness, abusive responses, or simply saying that they are too busy;
- the caller's ability to remain neutral, impartial and non-judgemental;
- how to record responses;
- how to end the interview.

It is also advisable, in order to avoid the frequent responses to 'cold-calling' (where the called person simply slams down the telephone), for the interviewer to contact the person in advance of the call, perhaps by mail, to indicate that the call will come, when, what it is about, and to ask for the recipient's cooperation in the project.

Many of the features of telephone interviewing are similar to those of effective interviewing per se, and we advise the reader to consult the comments on interviewing earlier and also in Chapter 25.

17.12 Comparing methods of data collection in surveys

Aldridge and Levine (2001, pp. 51–4) and Fowler (2009, pp. 80–3) offer useful summary guides to the advantages and disadvantages of several methods of data collection in surveys: personal face-to-face interviewing; telephone interviewing; self-administered/self-completion versus interviewer-administered; group administered; mailed surveys; delivered (distributed) surveys (e.g. personally delivered or delivered to an institution); Internet surveys. We refer the reader to these useful sources.

Additionally, Fowler (2009) and Dillman *et al*. (2014) discuss the benefits of combining methods of data collection (e.g. face-to-face interviews with telephone interviews, Internet surveys with postal surveys, advance

TABLE 17.3 ADVANTAGES AND DISADVANTAGES OF DATA-COLLECTION METHODS IN SURVEYS

	Advantages	*Disadvantages*	*Either advantages or disadvantages*
Postal	Time to think Costs may not be too expensive Opportunity for attractive survey design and graphics Complete at respondent's convenience, with opportunity for respondent to check Can reach many people No risk of interviewer intrusion or bias Can reach scattered populations Can gather sensitive data (nobody else is present) Can offer secure confidentiality, anonymity and non-traceability Standardized wording	Cost: printing, postage Time: response time and data entry Low response rates Need for contact details Risk of superficial coverage of topics No checking on understanding or seriousness of response Missing data Respondents may misunderstand instructions or items	Self-completion Impersonal Need for simple format Completion of sensitive information
Interviews face-to-face (individual)	Opportunity for gathering in-depth data Reduction of false responses Benefits of human-to-human contact and interpersonal behaviour High response rate Useful for exploring complex issues Opportunity to explain and clarify items and take questions from respondents Flexibility in item sequence Can build trust and rapport Ensure that only the respondent answers	Potential for perceived threat and bias in face-to-face meeting Costly: time for conducting interview, data entry and travel Not possible for large-scale survey Need to train interviewers Long data-collection period Access to sample Little time to think or reflect Flexibility can reduce standardization	Location of interviews Participation Personal Interviewer and interviewee characteristics Conduct of interview affects responses Small samples Standardization
Interviews (group)	Time-saving (compared to individual) Opportunity for gathering in-depth data Reduction of false responses Benefits of human-to-human contact and interpersonal behaviour High response rate Useful for exploring complex issues	Risk of 'group think' Potential for perceived threat and bias in face-to-face meeting Threat to confidentiality Not possible for large-scale survey Scheduling time and location for whole group to be present Costly: time for conducting interview, data entry Little time to think	Participation Personal Interviewer and interviewee characteristics Conduct of interview affects responses Small samples Standardization

	Advantages	*Disadvantages*	*Either advantages or disadvantages*
	Opportunity to explain and clarify items and take questions from respondents Flexibility in item sequence Can build trust and rapport Ensure that only the respondent answers		
Telephone	Honesty Anonymity (absence of the human face) Reduction in costs: time, money and travel Rapid contact Random dialling Access to dispersed sample and distant locations Response rate higher than postal survey Short data-collection period Opportunity to explain and clarify items Opportunity to probe participants Reduced interviewer and interviewee bias	Lack of visuals and non-verbal cues: oral and aural medium only Finding telephone numbers (particularly with cellphones) Easy for respondents to refuse or quit through the survey (i.e. to hang up) Limited time (no more than ten minutes) Cold calling has a bad name Time of day for calling may be inappropriate Biased sampling (no telephone) Respondents are ex-directory Immediacy: no time to think of responses Cost (phone charges) Personal answering the call may not be suitable Multiple-choice, rating scale and ranking questions are difficult Order effects can be strong Risk of socially desirable responses, satisficing and acquiescence	Sensitive questions: absence of an interviewer may encourage or discourage honesty of response Personal and yet impersonal Well-prepared and trained interviewer
Internet-based	Cost saving: time, money, data entry by researcher Speed: rapid distribution, completion and return Wide distribution: no problem of time and distance Access to minority and marginalized groups Opportunity for large samples and data volume Rapid data entry	Security of data and confidentiality Biased sampling (no Internet, or respondents' limited Internet expertise, volunteer samples) No checking on understanding or seriousness of response Need for email addresses or posting opportunities Multiple submissions Risk of superficial coverage of topics	Honesty of responses Impersonal Anonymity, confidentiality and privacy

continued

TABLE 17.3 CONTINUED

	Advantages	*Disadvantages*	*Either advantages or disadvantages*
	Easy access to people and dispersed populations Time to think Opportunity to complete it in stages, i.e. with time breaks Complete at respondent's convenience Opportunity for attractive survey design and graphics Higher response rates than postal surveys Environmentally friendly (no paper) Easy skip and branching arrangements Honest responses to sensitive issues Standardized wording Ease of data entry	Computer software compatibility and technical problems Limited number of items per screen Respondents give a minimal response Order effects People quit if it is too long or complex Missing data (or resentment if forced responses are required) Respondents regard it as spam Design expertise of the researcher Respondents may misunderstand instructions or items Overall low response rates Satisficing and acquiescence (see Chapter 18)	
Dropping off questionnaires	Opportunity to explain the survey face-to-face No training required for distribution staff (i.e. no interviews) Respondents have time to think and reflect Complete at respondent's convenience Higher response rates than postal survey	Costly: distribution staff and time	Impersonal

emails with interviews). Single mode surveys, write Dillman *et al*. (2014), are less effective than mixed mode surveys (e.g. telephone calls and emails, emails and websites, etc.) in terms of response rates.

Table 17.3 sets out advantages and disadvantages of these different types of survey administration.

We include Internet methods in Table 17.3, for ease of comparison with other methods, and our discussion turns to Internet surveys, devoting the next chapter entirely to this topic.

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Internet surveys

CHAPTER 18

This chapter will look at:

- the advantages and disadvantages of Internet surveys
- constructing Internet-based surveys
- ethical issues in Internet-based surveys
- sampling in Internet-based surveys
- improving response rates in Internet surveys
- technological advances

18.1 Introduction

The inclusion here of a chapter devoted to Internet surveys signals not only the prominence that these have in contemporary research but also raises key issues which Internet surveys highlight and which may not have such a high profile in more traditional surveys. Internet surveys bring a new perspective to existing issues in survey research, and we include these here. We advise readers to read this chapter in conjunction with Chapters 8, 17 and 24, and very many of the points in these three chapters apply convincingly to the present chapter.

Changes in the Internet and its uses are advancing rapidly, with access through a plethora of devices increasing exponentially. This chapter introduces issues which transcend particular mobile or computer devices and which can last over time. We recognize immediately that, in a chapter of this nature, huge technological changes will have occurred simply between the time of writing and the appearance of this book, and the future in this field has many unpredictable elements. The level of sophistication of mobile devices and their optimization for all kinds of research is advancing at breakneck speed. Could we even have imagined five years ago that email would become passé so quickly, or three years ago that SMS messaging would become yesterday's news, consigned to the older generation rather than the young digital natives of today whose life seems to revolve around apps and the absence of face-to-face communication (cf. Turkle, 2015)?

Internet surveys, whilst they have been the stuff of student evaluations of teaching for years (Morrison, 2013b), are becoming commonplace in many branches of educational research (Denscombe, 2014; Dillman *et al*., 2014; Roberts and Allen, 2015). Indeed they are becoming the predominant mode of conducting surveys, superseding paper-based surveys, be they through email or websites, on computers, cellphones, tablets and an ever-increasing range of electronic devices. Though they have much in common with paper-based surveys, they also have their own particular features, and we comment on these in this chapter.

Internet-based surveys can operate through, for example:

- an email with an introductory letter and questionnaire attachment (which requires the researcher to know the email address of the recipient);
- an email directing readers to a website where the survey can be found, or with a hyperlink to that site;
- an email with a survey embedded in it;
- a website which contains the survey: a web-based questionnaire;
- a general request for participation, placed in an electronic environment, such as an advertisement and connection to a survey placed on listservs, contact groups (e.g. special interest groups, newsgroups, discussion groups), blogs, social network sites and forums;
- a general advertisement;
- a dedicated website (one's own or others');
- public messages and advertising on social networking sites;
- companies who provide Internet survey services.

Internet surveys concern: the design of the survey (e.g. a questionnaire); its distribution and access to it; and data collection, storage and accessing data. We discuss these below.

18.2 Advantages of Internet surveys

An Internet survey claims several advantages in comparison to a paper survey:[1]

- *Costs*: it reduces costs (e.g. of postage, paper, printing, data entry by researchers, telephones, processing data, interviewer costs) and increases efficiency.

- *Speed*: it reduces the time taken to distribute, complete, gather and process data (data entered onto a web-based survey can be processed automatically as soon as the respondent enters the data rather than being keyed in later by the researcher), i.e. real-time data capture and processing.
- *Population and samples*: wider and much larger populations and samples can be accessed easily, enabling greater generalizability where Internet users come from a wide, diverse population. Internet surveys can yield representative data as they can reach a wider audience.
- *Contact*: e-surveys overcome spatial and temporal constraints (e.g. researchers and participants separated from each other in time, location and physical distance).
- *Volume*: a much larger volume of data can be collected.
- *Access*: researchers can reach populations that are otherwise difficult to reach (e.g. rare, minority, stigmatized, marginalized or deviant cases, as anonymity and non-traceability might be effective here).
- *Convenience*: respondents can complete the survey at a time to suit themselves, and they can complete the survey over time (i.e. they do not need to do it all at one sitting) and anywhere, i.e. in self-chosen and familiar settings.
- *Responses*: response rates may be higher (though some evidence challenges this). Participants tend to respond more quickly and more fully, reflectively and incisively, particularly when e-reminders are sent. Data are of a higher quality (richer, fuller responses with greater depth and reflection by participants). There are higher item completion and response, higher item variability and fewer missing values (though some evidence challenges this).
- *Ease*: it is easy to enter responses (e.g. click a radio button, check/tick a box).
- *Environment*: it is environmentally friendly (little or no paper).
- *Attractiveness*: graphics, different colours and fonts can be used.
- *Design flexibility*: skip-patterns and branching can be created and organized by the computer, so that participants do not have to understand complicated instructions, with automated, flexible navigation through the questionnaire consequent to the answers from respondents and boxes that they check/tick. It can enable diverse questions to be asked (through branching and skip functions).
- *Response checking*: the software can prompt respondents to complete missed/skipped items or to correct errors (e.g. two ticks for a single item in a rating scale), i.e. preventing them from continuing until a screen or item is completed: a 'forced response'. The computer can check incomplete or inconsistent replies.
- *Progress*: for each screen, an on-screen progress bar can show how much of the questionnaire has been completed (e.g. 50 per cent completed, 75 per cent completed).
- *Accuracy*: fewer missing or incorrect entries, as human error is reduced in entering and processing online data, these being done automatically.
- *Exportability*: data can be exported/imported into software (e.g. Excel, SPSS) for processing and subsequent analysis.
- *Anonymity, honesty and authenticity*: respondents may be more honest if their responses are anonymous and not face-to-face, i.e. a reduction in researcher effects, particularly if sensitive issues are being explored. Because of volunteered participation (i.e. an absence of coercion), greater authenticity of responses may be obtained.

With regard to costs, Watt (1997) notes that cost savings make a difference in comparison to a telephone survey, but that an Internet-based survey is only slightly less costly than a mail survey unless a web-based survey gathers data from more than around 500 participants, as the costs in terms of development and design time are considerable (though survey software has mitigated this). Fricker and Schonlau (2002) and Fox *et al*. (2003) suggest that the claims that Internet-based surveys are cheaper and faster are not always borne out by the evidence, and that, if Internet survey development, programming, testing and modification time, initial contact time and follow-up time to ensure an increased response rate are factored in, then the savings may not be as strong as the claims made. More recently, many Internet surveys have become less costly but not in terms of time for survey design and development.

18.3 Disadvantages of Internet surveys

An Internet survey also has several potential and actual disadvantages in comparison with a conventional survey (here the same references as those given above for 'advantages of Internet surveys' also give disadvantages, see note 1), for example:

- *Spam*: recipients and servers may regard the request to participate as spam or junk mail, and ignore or delete it.

- *Expertise*: respondents may not have sufficient Internet expertise to complete it correctly.
- *Sampling*: some target groups may not have access to the Internet and some respondents/groups may not be available or may not answer (non-response), particularly if the questions are sensitive. There may be sampling bias/skew, as the researcher has less control over the population and sampling, and respondents self-select. Random sampling may be impossible (which may affect statistics used, i.e. preventing use of some parametric statistics). Personal details may be withheld or unverifiable, hence sub-sampling and parametric analysis may be impossible.
- *Abandonment and dropout*: it is easy for respondents to simply stop altogether, or to send incomplete surveys, or not to send them at all. This is particularly the case if surveys are long or if researchers try to coerce respondents into completion by the use of a forced response (making them unable to continue until every item has been completed on a single screen). As a general rule, Internet surveys should be shorter than paper-based surveys or non-response can occur.
- *Computer difficulties*: the different configuration and software of computers may affect access to the survey, its layout, presentation and speed of connection. The computer or server may 'crash', 'freeze' or refuse to work, particularly if the survey is long or contains a lot of graphics. Error messages may appear. The survey may require software that respondents do not have or which downloads very slowly. Servers, computer networks, connections or the 'system' may go down. Surveys embedded in emails should be short, with few or no graphics or embedded objects, as the recipients' computer may not receive these.
- *No interviewer*: this is an impersonal medium, with no opportunities for in-depth probes.
- *Security, privacy and confidentiality*: these are not absolute, as hackers, data miners and spammers may intercept and track participants, and results are stored by software providers. Emails are particularly vulnerable.
- *Design matters*: technical expertise is needed if the researcher is going to write and design her/his own survey, i.e. not using commercially produced, open access or free software.
- *Instructions and answering*: if instructions are misunderstood, unclear, too complicated or too many, this may lead to irrelevant, unreliable or no responses, and if respondents are unclear how to answer then they are likely to simply skip items or stop.
- *Time*: surveys embedded in emails may require subsequent manual data entry by the researcher.
- *Response rates*: these tend to be low in Internet surveys (calculated in relation to those who accessed the survey). Response rates drop off if the survey is long or too complicated.
- *Misreporting*: there may be fake and false reporting: respondents may give deliberately fake, incorrect or socially desirable responses, or complete the survey to obtain a promised reward. There may be multiple submissions from the same person. Verifying the identity or details of participants may be impossible.
- *Satisficing*: respondents, particularly in a long survey, will enter any response rather than no response, compromising the quality, accuracy and reliability of the response (discussed below). The quality of responses becomes increasingly questionable as a long survey progresses.

Several of the positive claims made for Internet surveys are questionable. For example, Fricker and Schonlau (2002) and Fox *et al*. (2003) note that, in comparison to conventional surveys: (a) response rates may not be higher; (b) time taken in the several stages of the survey – from design to data collection and analysis – may not be shorter, even though delivery time is shorter; (c) data quality may be no better: coverage of items, sampling, response and non-response, honesty; and (d) potential respondents may not have access to the Internet and may be computer illiterate.

18.4 Constructing Internet-based surveys

Much Internet survey design can use any of the many online Internet survey templates and (often free) services available, and the companion website provides websites which provide such services. At the time of writing, widely used providers include: SurveyMonkey; Zoomerang; Free Online Surveys; SmartSurvey; Survey Planet; Google Forms. We provide the websites of these and others on the companion website. Several of these also automatically collate and present results so that they can be downloaded into, for example, Excel, SPSS, SAS or STATA.

For researchers wishing to construct their own Internet surveys with software and templates, there are several points of guidance below. Internet surveys can be made more attractive than their paper-based counterparts, with graphics, fonts, colours etc., but there are cautions, and we indicate these below.

Dillman *et al*. (1998a, 1998b, 1999, 2014) set out several concerns for Internet surveys, some technical

and some presentational. In terms of technical matters, they found that the difference between simple and 'fancy' versions of questionnaires (the former with few graphics, the latter with many, using sophisticated software) could be as much as three times the size of the file to be downloaded, thereby increasing download time. Respondents with slow browsers or limited power either spent longer in downloading the file or, indeed, the machine crashed before the file was downloaded. Download speeds continue to vary in different parts of the world and at different times of the day. They also found that recipients of plain versions were more likely to complete a questionnaire than those receiving fancy versions, as it took less time to complete the plain version. Utilizing advanced page layout features does not translate into higher completion rates, indeed more advanced page layout reduced completion rates. Similarly, Fricker and Schonlau (2002) report studies indicating a 43 per cent response rate to an email survey compared to a 71 per cent response rate for the same mailed paper questionnaire, and that higher response rates in an Internet survey are typically only obtained from specialized samples (e.g. undergraduates).

For presentational matters, Dillman *et al*. (1998a, 1999, 2014) comment that in a paper-based survey the eyes and the hands focus on the same area, whilst in an Internet survey the eyes focus on the screen whilst the hands often focus on the keyboard or the mouse, rendering completion more difficult. This is one reason to avoid asking respondents to type in many responses to open-ended questions, and to replace this with radio buttons or clicking a check box. The researchers also found that 'check-all-that-apply' lists of factors had questionable reliability, as respondents tended to complete those items at the top of the list and ignore the rest. Hence they recommend avoiding the use of large check-all-that-apply lists in a web-based survey.

It is important to keep the introduction to the questionnaire short (no more than one screen) and informative (e.g. about how to move on) and to avoid giving a long list of instructions. Further, as the first question in a survey tends to raise in respondents' minds a particular mindset, care is needed in setting the first question, to entice participants and not to put them off participating (e.g. not too difficult, not too easy, interesting, straightforward to complete, avoiding drop-down boxes and scrolling). Dillman *et al*. (1998a, 1998b, 1999, 2014) make recommendations about the layout of the screen, for example, keeping the response categories close to the question for ease of following; using features like brightness, large fonts and spacing for clarity; and avoiding too many changes of font and font size. They also suggest following the natural movement of the eyes from the top-left (the most important part of the screen, hence the part in which the question is located) to the bottom-right quadrants of the screen (the least important part of the screen which might contain the researcher's logo). They comment that the natural movement of the eye is to read prose unevenly (e.g. saccadically), with the risk of missing critical words, particularly on long lines; hence they advocate keeping lines and sentences short. It is also useful to include a progress bar to indicate how much of the questionnaire has been completed.

Some respondents may have less developed computer skills than others, and may not be familiar with web-based questionnaires, for example, with radio buttons, scroll bars, drop-down menus, where to insert open-ended responses. Hence the survey designer must not overestimate the capability of the respondent to use the software. Indeed explanations on how to respond may have to be outlined in the survey itself.

Dillman *et al*. (1999, 2014) suggest that the problem of differential expertise in computer usage can be addressed in three ways:

1. Place the instructions for how to complete the item next to the item itself (not all placed together at the start of the questionnaire).
2. Ask the respondents at the beginning about their level of computer expertise; if they are more expert, offer them the questionnaire with certain instructions omitted, and if they are less experienced, direct them to instructions and further assistance.
3. Have a minimized 'floating window' that accompanies each screen and which can be maximized to give further instructions.

Some web-based surveys prevent respondents from proceeding until they have completed all the items on the screen (a forced response). Whilst this might ensure coverage, it can also anger respondents – such that they give up and abandon the survey. Some web-based surveys prevent respondents from having a deliberate non-response (e.g. if they do not wish to reveal particular information, or if, in fact, the question does not apply to them, or if they do not know the answer). The advice of Dillman *et al*. (1999, 2014) is generally to avoid forced responses and to give respondents categories of 'prefer not to answer'/'decline to answer', 'don't know', 'not applicable' and 'other'. It is much easier for participants in a web-based survey to abandon the survey – a simple click of a button – so more attention has to be given to keeping them participating than in a paper-based survey.

The location of the instruction (e.g. to the right of the item, underneath the item, to the right of the answer box) is important. Locating the instruction too far to the right of the answer box (e.g. more than nine characters of text to the right) can mean that it is outside the foveal view (two degrees) of the respondent's vision, and hence can be overlooked. Redline *et al*. (2002) advocate making an instruction easier to detect by locating it within the natural field of vision of the reader, setting it in a large font to make it bolder and using a different colour. If the researcher wishes to include skips and branches then this can be done automatically 'behind the scenes', i.e. if the participant gives a particular response then the computer automatically takes him or her past a skipped part or to a branching part.

Redline *et al*. (2002) identify many other variables that impact on the success of Internet surveys:

- the greater the number of words, the more the reader will be absorbed with the question than with the instructions;
- if there are more than seven response categories per item the reader may make errors;
- response-order effects: respondents in a self-administered survey tend to choose earlier rather than later items in a list (the primacy effect);
- if respondents are asked to write an open-ended response they may overlook instructions, as they are absorbed in composing their own response, and the instruction may be out of their field of vision when writing their answer;
- items located at the bottom of a page are more likely to elicit a non-response than items further up a page, and instructions near the bottom of a page are more likely to be overlooked, so avoid this;
- if instructions are located too far from the answer box they may be overlooked.

This advice applies not only to online survey questionnaires but also to paper-based surveys.

There are several 'principles' for designing web-based questionnaires (e.g. Schaefer and Dillman, 1998; Dillman *et al*., 1999; Dillman and Bowker, 2000, pp. 10–11; Shropshire *et al*., 2009; Dillman *et al*., 2014):

- Consider the capabilities and configurations of the respondents' computers and the respondents themselves.
- Ensure that the layout/presentation of the survey will be the same across platforms, servers, browsers and respondents, and avoid differences in the visual appearance of questions that may happen as a result of different computers, configurations, operating systems, screen displays (e.g. partial and wrap-around text) and browsers.
- Enable the survey to run on computers, cellphones, iPads and other different devices.
- Ensure that security, confidentiality and privacy are in place.
- Start the web questionnaire with an interesting welcome screen that motivates respondents to continue, makes it clear that it is easy to complete, gives clear instructions on how to proceed and contains information and a check box for informed consent.
- Provide a PIN number if you wish to limit access to those people in the sample.
- Ensure that the first question can be seen in its entirety on the first screen, and is easy to understand and complete.
- Embed visual images in a survey and place interest-based questions early, as this reduces premature dropout from the survey.
- Ensure that the layout of each question is similar to a paper format, as respondents may be familiar with this.
- Ensure that the use of colour keeps figure/ground consistency and readability, so that it is easy to navigate through the questionnaire with navigational flow unimpeded, and so that the measurements used in questions are clear and sustained.
- Ensure consistency in colours, fonts, layout.
- Keep the line length short, to fit screen size and respondent focus.
- Minimize the use of drop-down boxes and direct respondents to them where they occur.
- Give clear instructions for how to move through the questionnaire using the computer, and keep instructions for computer actions at the point where the action is needed, rather than placing them all at the start of the questionnaire.
- Avoid forced responses (requiring respondents to answer each question before being able to move on to the next question/screen).
- Ensure that questionnaires scroll easily from question to question, unless order effects are important.
- If multiple choices are presented, keep them to a single screen; if this is not possible then consider double columns, with navigational instructions.
- Consider providing a progress bar to indicate how far the respondent has reached in the survey (it may or may not be advisable if the questionnaire is quite long, hence judgement is required).
- Avoid tick-all-those-that-apply kinds of question.
- Enable respondents to save their survey and complete it later, and to keep their own copy (backup) of their completed survey.

- Have a 'thank you' screen for when the respondent submits the completed survey.
- Provide a 'help' button for further explanation and contact details of the researcher.

Additionally, Heerwegh *et al*. (2005) report that personalizing the survey (i.e. using the recipient's name in the salutation) increases response rates by 8.6 per cent (e.g. starting a survey letter with Dear [name of specific person]), though care has to be taken to ensure consistency, for example, it is counter-productive to start an Internet survey letter with Dear [name of person] and then, later, to refer to 'student', 'colleague' etc. (i.e. a depersonalized version). Personalizing a survey increases the chances of a respondent *starting* the survey (p. 92) rather than dropping out during the survey (i.e. they might still drop out later). Whether participants continue and complete the survey depends on other factors, for example: difficulty in completing the items; relevance of the topic to the respondents; and user-friendliness of the survey. They also report (p. 94) that personalizing an e-survey increases participant honesty on sensitive matters (e.g. number of sexual partners), increases adherence to survey instructions (p. 96) and increases the tendency of respondents to answer questions in a socially desirable way.

Denscombe (2009b, pp. 286–7) reports that, for online surveys, fixed-choice questions tend to have a 'lower item-response rate than open-ended questions', and, for open-ended questions, item non-response is lower in online surveys than in paper-based surveys.

Christian *et al*. (2009) report that if a survey questionnaire presents the positive end of a scale first then, whilst it may not make a difference to the response, it does increase response time (which might lead to dropping out). They also note that if the positive categories are given lower numbers (e.g. '1' and '2') and the negative categories are given high numbers (e.g. '4' and '5'), then this can increase response time. Further, they report that displaying categories in several columns increases response time, as does giving the poles of rating scales numbers rather than words.

Mora (2011b) notes that more positive results are obtained if the scale runs from –2 to +2 or from –4 to +4, that scales with the positive label on the left and higher numerical ratings on the left had significantly higher ratings, that scales with 'definitely agree' on the left and 'definitely disagree' on the right tended to have higher agreements, and higher scores were recorded for positively worded items. The percentage of affirmative responses was higher in a paper-based survey than in an Internet-based survey (Dillman *et al*., 2003; Morrison, 2013b).

Toepel *et al*. (2009) suggest that account has to be taken of the 'cognitive sophistication' of the respondents, as those with less cognitive sophistication tend to be affected by contextual clues more than those with more cognitive sophistication. Context effects also occur when a particular item is affected by the items around it or which precede it, in effect providing cues for the respondent, or in which a particular mindset of responses is created in the respondent (Friedman and Amoo, 1999).

The importance of the visual aspect of questionnaires is heightened in Internet surveys (Smyth *et al*., 2004), and this concerns the layout of questions, instructions and response lists, the grouping of items, the colours used, the spacing of response categories and the formatting of responses (e.g. writing in words or checking boxes). Smyth *et al*. report that respondents use 'preattentive processing' when approaching Internet surveys, i.e. they try to take in and understand the whole scene (or screen) before attending to specific items, hence visual features are important, for example, emboldened words, large fonts, colours, brightness, section headings, spacing, placing boxes around items. This rests on Gestalt psychology which abides by the principles of: (a) *proximity* (we tend to group together those items that are physically close to each other); (b) *similarity* (we tend to group together those items that appear alike); (c) *prägnanz* (figures or items with simplicity, regularity and symmetry are more easily perceived and remembered).

Smyth *et al*. (2004) also suggest (p. 21) that headings and separation of sections take on added significance in Internet-based surveys. Separating items into two sections with headings had a 'dramatic effect' on responses, as respondents felt compelled to answer both sub-groups (70 per cent gave an answer in both sub-groups whereas only 41 per cent did so when there were no headings or sectionalization). They also found that separating a vertical list of items into sub-groups and double columns should be avoided. They report that asking for open-ended responses (e.g. writing their subject specialisms) can be more efficient than having them track down a long list of subjects (e.g. from a drop-down menu) to find the one that applies to them, though this can be mitigated by placing simple lists in alphabetical order. Finally, they found that placing very short guides underneath the write-in box rather than at its side (e.g. dd/mm/yy for 'day/month/year') increased response rates, and that placing instructions very close to the answer box improved response rates.

Dillman *et al*. (2003) also found that having respondents use a 'yes/no' format and having a 'forced choice' (no option but to answer) for responding

resulted in increased numbers of affirmative answers, even though this requires more cognitive processing than non-forced choice questions (e.g. 'tick[check]-all-that-apply' questions) (p. 23). This is because respondents may not wish to answer questions in the outright negative (p. 10). Even if they do not really have an opinion, or they are neutral, or the item does not really apply to them, they may choose a 'yes' rather than a 'no' category. They may leave a blank rather than indicating a 'no'.

Dillman *et al*. (2003) report that respondents tend to select items higher up a list than lower down a list of options (the primacy effect), and opt for the 'satisficing' principle (they are satisfied with a minimum sufficient response, selecting the first reasonable response in a list and then moving on rather than working their way down the list to find the optimal response), i.e. item order is a significant feature, making a difference to responses of over 39 per cent (p. 7). This is particularly so when respondents are asked for opinions and beliefs rather than topics seeking factual information. They also suggest that the more difficult the item is, the more respondents will move towards 'satisficing'. 'Satisficing' and the primacy effect were stronger in Internet surveys than paper-based surveys (p. 22), and changing 'check-all-that-apply' to forced responses ('yes/no') did not eliminate response order effects. Similarly, Diaz de Rada and Dominguez (2015) report that participants change their minds about participation as the survey proceeds, but, rather than withdrawing altogether, move towards satisficing, giving answers that require minimum effort rather than thinking deeply, and that they give affirmative answers in the spirit of acquiescence or 'don't know' answers.

Dillman *et al*. (2003) report that the order of response categories can have an effect on responses, citing a study that found that asking college students whether their male or female teachers were more empathetic was affected by whether the 'male' option was placed before or after the 'female' option: female teachers were evaluated more positively when respondents were asked to compare them to male teachers than when male teachers were compared to female teachers (p. 6). Respondents compare the second item in light of the first item in a list rather than considering the items independently.

Internet-based surveys, then, raise several challenges and problems. Some of these are indicated in Table 18.1, together with possible solutions.

18.5 Ethical issues in Internet-based surveys

In Internet surveys, issues of informed consent, anonymity, privacy and confidentiality, non-traceability, protection from harm, the precautionary principle and data security are important (e.g. Fox *et al*., 2003; Hammersley and Traianou, 2012). Chapter 8 addressed ethical issues in online research and these apply to Internet surveys. Ethics in Internet surveys also concerns the offering of incentives to participate, to ensure that they are appropriate and not excessive. Marshall and Rossman (2016) suggest that researchers must consider several ethical issues in researching online communities (pp. 182–3):

- how public and private the data are;
- the sensitivity of the topic;
- the degree of interaction between the researcher and the participants and between participants;
- the vulnerability of the participants.

Given that the URL identifies a specific location and is held on servers, and given that some survey providers have control over this, it is impossible to guarantee total security of identity and information here. Rather, steps have to be taken to protect data in terms of collection, storage, access and electronic transfer (just as with non-Internet surveys).

To address security and confidentiality of identity, the researcher may require no identification details from the participant and ask for none in the survey. Or the researcher may take steps to protect data privacy, anonymity and confidentiality, through password protection, one-way scrambling of the machine number when the data are being submitted ('encoding it in a manner that cannot subsequently be decoded'; Fox *et al*., 2003, p. 178). However, this does not give any absolute guarantee that hackers, data miners or spammers will not intercept and access the data, particularly where it can be linked to an email address. An email survey can reveal the identity and traceability of the respondent. Security (e.g. through passwords and PIN numbers) is one possible solution, though this, too, can create problems as respondents may feel that they are being identified and tracked, and indeed some surveys may deposit unwelcome 'cookies' onto the respondent's computer, for future contact.

Completing a survey may be taken to indicate informed consent, and Roberts and Allen (2015) note the American Educational Research Association's (2000) Code of Ethics which indicates (Clause 13.01.b)

TABLE 18.1 PROBLEMS AND SOLUTIONS IN INTERNET-BASED SURVEYS

Problem (sampling)	*Possible solution*
Some subsample groups may be under-represented in the respondents	Adjust the results by weighting the sample responses (see the comments on a 'boosted sample' and 'weighting' in Chapter 12)
There may be coverage error (not everyone has a non-zero chance of being included)	Disclose the sample characteristics in reporting
Non-response and volunteer bias	Follow-up messages posted on websites and electronic discussion groups. Use emails to contact potential participants. Require the respondents to submit their replies screen by screen. (This enables the researcher not only to use some data from incomplete responses, but also enables her to identify in detail patterns of non-response, i.e. responding is not an all-or-nothing affair (either submit the whole questionnaire or none of it) but can be partial (a respondent may answer some questions but not others))
Problem (ethics)	*Possible solution*
Respondents may wish to keep their identity from the researcher, and an email address identifies the respondent (in the case of sensitive research, e.g. on child abuse or drug abuse, this may involve criminal proceedings if the identity of the respondent is known or able to be tracked by criminal investigators who break into the site). Non-traceability of respondents may be problematic	Direct respondents to a website rather than to using email correspondence. Provide advice on using non-traceable connections to access and return the survey (e.g. an internet café, a library, a university). Advise the respondent to print off the survey and return it by post to a given address. Avoid asking respondents to enter a password or to give an email address. Prevent access to unprotected directories and confidential data
Respondents may not know anything about the researcher, or if it is a *bona fide* piece of research and not simply a marketing ploy	Include the researcher's affiliation (e.g. university), with a logo if possible
Informed consent	Ensure that it is easy for respondents to withdraw at any time (e.g. include a 'Withdraw' button at the foot of each screen)
Problem (technical: hardware and software)	*Possible solution*
The configuration of the questionnaire may vary from one machine to another (because of web browsers, connection, hardware, software) and can lead to dropout	Opt for simplicity. Test the survey on different computer systems/browsers to ensure consistency. Avoid surveys that require real-time completion
The screen as set out by the survey designer may not appear the same as that which appears on the respondent's screen	Opt for simplicity. Use a commercial survey software system for generating the questionnaire. Avoid high level programs
Slow network connections or limited bandwidth can slow down loading	Keep the use of graphics to a minimum. Advise on the possible time it takes to load
Respondents may not have the same software, or the same version of the software as the sender, rendering downloading of the questionnaire either impossible or distorting the received graphics	Avoid the use of graphics and more advanced software programs

Problem (technical: hardware and software)	*Possible solution*
Graphics may be corrupted/incompatible between the sender and the user, i.e. between one kind of machine, user platform and software and another. Hardware may differ between sender and receiver	Opt for simplicity. Use commercially available web-based surveying systems and packages. Use image files (e.g. .jpeg, .gif) to reduce loading time. Avoid pop-ups if possible as they reduce response rate
The greater the use of graphics and plug-ins (e.g. using Java and Applets), the longer it takes to download, and, particularly – though not exclusively – if respondents do not have broadband access then time-consuming downloads could result in either the respondent giving up and cancelling the download, or creating a bad mood in the respondent	Keep software requirements as low-tech as possible. Avoid questionnaires that use sophisticated computer graphics
There may be slow loading times due to internet congestion	Avoid sophisticated graphics and 'fancy' presentations as these take longer to download
The physical distance between points on an attitude scale may spread out because of configuration differences between machines	Indicate how best the questionnaire may be viewed (e.g. 800 × 400)
The construction procedures for wrap-around text may vary between computers	Keep lines of text short
Email questionnaires may distort the layout of the questionnaire (some email software uses HTML, others do not)	Avoid sending a questionnaire directly using email; rather, post it on a website (e.g. so that respondents visit a website and then click a box for immediate transfer to the questionnaire). Consider using an email to direct participants to a website (e.g. the email includes the website which can be reached by clicking in the address contained in the email). Use an email that includes an attachment which contains the more graphically sophisticated survey instrument itself
Problem (respondents)	*Possible solution*
Respondents may be unfamiliar or inexperienced with the internet and the media	Keep the questionnaire simple and easy to complete
Respondents may send multiple copies of their completed questionnaire from the same or different addresses	Have a security device that tracks and limits (as far as possible) respondents who may be returning the same questionnaire on more than one occasion. Use passwords (though this, itself, may create problems of identifiability). Collect personal identification items. Check for internal consistency across submissions
There may be more than one respondent to a single questionnaire (the same problem as in, for example, a postal questionnaire)	Include questions to cross-check the consistency of replies to similar items
Respondents may not be used to pull-down menus	Provide clear instructions
Drop-down boxes take up more space on a screen than conventional questionnaires	Avoid their overuse
Respondents dislike the situation where the computer prevents them from continuing to the next screen until all the items on a particular screen have been completed	Avoid this unless considered absolutely necessary

continued

TABLE 18.1 CONTINUED

Problem (respondents)	*Possible solution*
The language of email surveys can risk offending potential participants ('flaming')	Check the language used to avoid angering the participants
Respondents' difficulty in navigating the pages of the online survey	Keep instructions to the page in question. Make the instructions for branching very clear (font size, colour, etc.)
Problem (layout and presentation)	*Possible solution*
A page of paper is longer than it is wide, but a screen is wider than it is long, and a screen is smaller than a page, i.e. layout becomes a matter of concern	Remember that screen-based surveys take a greater number of screens than their equivalent number of pages in a paper copy. Sectionalize the questionnaire so that each section fills the screen, and does not take more than one screen
The layout of the text and instructions assumes greater importance than for paper questionnaires	Opt for clarity and simplicity
The layout uses a lot of grids and matrices	Avoid grids and matrices: they are a major source of non-response
The order of items affects response rates	Locate requests for personal information at the beginning of the survey. Include 'warm-ups' and early 'high hurdles' to avoid dropout
Respondents may be bombarded with too much information in an introductory message	Place the advertisement for the survey on user groups as well as for the general public, inviting participants to contact such-and-such a person or website for further information and the questionnaire itself, i.e. separate the questionnaire from the advertisement for/introduction to the questionnaire
Respondents may be overloaded with instructions at the beginning of the survey	Avoid placing all the instructions at the start of the questionnaire, but keep specific instructions for specific questions
Respondents may be overloaded with information at the beginning of the survey	Keep the initial information brief and embed further information deeper in the survey
Respondents may have to take multiple actions in order to answer each question (e.g. clicking on an answer, moving the scroll bar, clicking for the next screen, clicking to submit a screen of information)	Keep the number of actions required in order to move on to a minimum
Respondents may not be able to see all the option choices without scrolling down the screen	Ensure that the whole item and options are contained on a single screen
Respondents may not understand instructions	Provide a helpline, email address or contact details of the researcher. Pilot the instrument
Instructions about options may be unclear	Use radio buttons for single choice items, and try to keep layout similar to a paper layout
Respondents only read part of each question before going to the response category	Keep instructions and words to a necessary minimum

Problem (reliability)	*Possible solution*
Respondents may alter the instrument itself. The researcher relinquishes a greater amount of control to the respondents than in conventional questionnaires	Include technological safeguards to prevent alteration and have procedures to identify altered instruments
Respondents may be forced to answer every question even when they consider some response categories inappropriate	Pilot the survey. Include options such as 'don't know' and 'do not wish to answer' and avoid forcing respondents to reply before they can move on
Respondents may not be telling the truth – they may misrepresent themselves	Include questions to cross-check replies (to try to reduce the problem of respondents not telling the truth)
Problem (dropout)	*Possible solution*
Respondents may lose interest after a while and abandon the survey, thereby losing all the survey data	Have a device that requires respondents to send their replies screen by screen (e.g. a 'Submit' button at the foot of each screen), section by section, or item by item. Put each question or each section on a separate screen, with 'submit' at the end of each screen. Adopt a 'one-item-one-screen' technique
Respondents may not know how long the questionnaire is, and so may lose interest	Include a device for indicating how far through the questionnaire the respondent has reached: a progress bar at the bottom or the side of the survey
Internet surveys take longer to complete than paper-based surveys	Keep the internet survey as short, clear and easy to complete as possible
People do not want to take part, and it is easier for someone to quit or cancel an internet-based survey than a paper-based survey (simply a click of a button)	Increase incentives to participate (e.g. financial incentives, lottery tickets (if they are permitted in the country))
Diminishing returns (the survey response drops off quite quickly). Newsgroup postings and electronic discussion group data are removed, relegated or archived after a period of time (e.g. a week), and readers do not read lower down the lists of postings	Ensure that the website is re-posted each week during the data collection period
Non-participation may be high (i.e. potential participants may not choose to start, in contrast to those who start and who subsequently drop out)	Increase incentives to participate. Locate personal informational questions at the start of the survey
Error messages (e.g. if an item has not been completed) cause frustration and may cause respondents to abandon the questionnaire	Avoid error messages if possible, but, if not possible, provide clear reasons why the error was made and how to rectify it

that waivers of consent may apply to Internet surveys if there is minimal risk or if the obtaining of consent may be impractical. Nevertheless this does not exonerate the researcher from indicating to participants that, whilst every reasonable step has been taken to protect confidentiality and to prevent unwanted access, it cannot be a watertight guarantee, even if this means that some respondents decline to participate. Transparency and honesty trump the researcher's personal wishes for responses here.

Researchers can provide initial information about ethical matters at the front of the survey, with rights: (a) to withdraw; (b) not to answer specific questions; and (c) to withdraw their consent freely. Researchers can provide a check box for participants to give consent, and indicate to what they are giving consent. Such consent, Roberts and Allen (2015) aver, should be free of coercion, and care should be taken to ensure that this is the case for potentially vulnerable participants (e.g. those with special needs, minorities, children etc.).

18.6 Sampling in Internet-based surveys

Watt (1997) suggests that there are three types of Internet sample:

- an *unrestricted* sample (anyone can complete the questionnaire, but it may have limited representativeness);
- a *screened* sample (quotas are placed on the subsample categories and types, e.g. gender, income, job responsibility etc.);
- a *recruited* sample: respondents complete a preliminary classification questionnaire and then, based on the data received, are recruited or not.

Regardless of sampling type, sampling bias is a major concern for Internet-based surveys (Coomber, 1997; Roztocki and Lahri, 2002; Schonlau *et al*., 2009), for example, 'sampling representativeness and validity of data' (Hewson *et al*., 2003, p. 27). The view of over-representation of some and under-representation of others is challenged (Hewson *et al*., 2003) by results showing that samples taken from users and non-users of the Internet did not differ in terms of income, education, sexual orientation, marital status, ethnicity and religious belief. Nonetheless, they did differ in terms of age, with Internet samples containing a wider age range than non-Internet samples, and in terms of sex, with Internet samples containing more males. Hewson *et al*. report overall a greater diversity of sample characteristics in Internet-based samples, though they caution that this is inconclusive, and that the characteristics of Internet samples, like non-Internet samples, depend on the sampling strategy used. Stewart and Yalonis (2001) suggest that one can overcome the possible bias in sampling through simple stratification techniques.

A problem in sampling in Internet surveys is estimating the size and nature of the population from which the sample is drawn, a key feature of sampling. Researchers may have no clear knowledge of the population characteristics or size. The number of Internet users is not a simple function of the number of computers or the number of servers (e.g. many users can employ a single computer, cellphone, iPad or server, and many users have all of these or more than one smartphone). Further, it is difficult to know how many or what kind of people see a particular survey on a website (e.g. more males than females), i.e. the sampling frame is unclear. Moreover, certain sectors of the population may be excluded from the Internet, for example, those not wishing to or unable (e.g. because of cost or availability or ability) to have access to the Internet.

Internet-based surveys are based largely on volunteer samples (see Chapter 12), obtained through general posting on the web (e.g. an advertisement giving details and directing volunteers to a site for further information), or, for example, through announcements to specific newsgroups and interest and user groups on the web, for example, SchoolNet. Lists of different kinds of user (USENET) groups, newsgroups and electronic discussion groups (e.g. listservs) can be found on the web. Several search engines exist that seek and return web mailing lists, such as: www.liszt.com (categorized by subject); Catalist (the official catalogue of listserv lists at www.lsoft.com/catalist.html); and Meta-List.net (www.meta-list.net), which searches a database of nearly a quarter-of-a-million mailing lists. Dochartaigh (2002) and Denscombe (2014) provide useful material on web searching for researchers.

In Internet surveys the researcher is using non-probability, volunteer sampling, and this may decrease the generalizability of the findings (this may be no more a problem on Internet-based surveys than on other surveys). Opportunity samples (e.g. of students, or of particular groups) may restrict the generalizability of the research, but this may be no worse than in conventional research, and may not be a problem so long as it is acknowledged. Volunteers may differ from non-volunteers in terms of personality (e.g. they may be more extravert or concerned for self-actualization (Bargh *et al*., 2002)) and may select themselves into, or out of, a survey, again restricting the generalizability of the results.

One method to try to overcome the problem of volunteer bias is to strive for very large samples, or to record the number of hits on a website, though these are crude indices. Another method of securing the participation of non-volunteers in an Internet survey is to contact them by email (assuming that their email addresses are known), for example, a class of students, a group of teachers. However, email addresses themselves do not give the researcher any indication of the sample characteristics (e.g. age, sex, nationality etc.). Gwartney (2007, p. 17) suggests that online surveys might be most appropriate with 'closed populations', i.e. employees in a particular organization, as this will enable the researcher to know some of the characteristics and parameters of the respondents.

18.7 Improving response rates in Internet surveys

Despite mobile phone optimization and increasing access to, and country-wide penetration of, the Internet, response rates for Internet surveys are typically lower

than for a paper-based survey and their equivalent mail surveys (Solomon, 2001), as is the rate of completion of the whole survey (Witmer *et al*., 1999; Reips, 2002a; Morrison, 2013b). Witmer *et al*. (1999) found that short versions of an Internet-based questionnaire did not produce a significantly higher response rate than the long version (p. 155). Solomon (2001) suggests that response rates can be improved through the use of personalized email, follow-up reminders, using simple formats and pre-notification of the intent to survey.

Reips (2002a) provides useful guidelines for increasing response rates on an Internet survey, for example, by having several websites and postings on several discussion groups that link potential participants or web surfers to the website containing the questionnaire. He also suggests utilizing a 'high-hurdle technique', where 'motivationally adverse factors are announced or concentrated as close to the beginning' as possible (p. 249), so that any potential dropouts self-select at the start rather than during the data collection. A 'high-hurdle' technique, he suggests, comprises:

- *seriousness*: inform the participants that the research is serious and rigorous;
- *personalization*: ask for an email address or contact details and personal information;
- *impression of control*: inform participants that their identity is traceable;
- *patience: loading time* – use image files to reduce loading time of web pages;
- *patience: long texts* – place most of the text in the first page, and successively reduce the amount on each subsequent page;
- *duration*: inform participants how long the survey will take;
- *privacy*: inform the participants that some personal information will be sought;
- *preconditions*: indicate the requirements for particular software;
- *technical pre-tests*: conduct tests of compatibility of software;
- *rewards*: indicate that any rewards/incentives are contingent on full completion of the survey.

Some of these strategies could backfire on the researcher (e.g. the disclosure of personal and traceable details), but the principle here is that it is often better for the participant not to take part in the first place rather than to drop out during the process. (Frick *et al*. (1999) found that early dropout was not increased by asking for personal information at the beginning.)

Reips (2002a) also advocates the use of 'warm-up' techniques in Internet-based research in conjunction with the 'high-hurdle' technique. He suggests that most dropouts occur earlier rather than later in data collection, or indeed at the very beginning (non-participation), and that most such initial dropouts occur because participants are overloaded with information early on. He suggests that it is preferable to introduce some simple-to-complete items earlier on to build up an idea of how to respond to the later items and to try out practice materials.

Researchers can try several ways to improve response rates (e.g. Schaefer and Dillman, 1998; Frick *et al*., 1999; Crawford *et al*., 2001; Dillman *et al*., 2009, 2014; Mora, 2010; Monroe, 2012; Morrison, 2013b; Denscombe, 2014):

- send an advance introductory letter by email, indicating the purposes, contents and time needed for the survey;
- consider incentives (e.g. a lucky draw, payment);
- have a welcome screen, which includes the institution (and its logo) and messages of support from senior people;
- make the instructions and questions clear and easy to answer;
- avoid asking for unnecessary identifying information (e.g. email addresses);
- keep it short (taking no more than 10–15 minutes (at most) to complete);
- keep the design consistent, clear, uncomplicated, attractive and easy to understand;
- send follow-up reminders (a maximum of three reminders);
- state anonymity and non-traceability (it may be impossible to offer 100 per cent guarantees here, but researchers can state that all steps have been taken to address these);
- personalize the survey: 'Dear ____' [name];
- avoid 'forced responses' if possible, and include the options (where relevant) of 'decline to answer', 'don't know', 'not applicable' or 'other';
- avoid items which require too much respondent effort (e.g. too much memory recall, comprehension, knowledge of technical vocabulary, concepts);
- avoid wordy questions, unclear concepts or asking more than one thing in the same question.

Response rates can also be improved through ease of question formats and ease of answering, for example, with check boxes or radio buttons for:

- yes/no questions;
- multiple-choice questions (select one by using a radio button);

- multiple-choice questions (select an exact number or as many as you wish);
- a matrix of multiple questions with the same response scales (e.g. rating scales);
- horizontal scales (preferable to vertical scale);
- drop-down lists with a single choice.

Further, make it easy to enter responses for continuous rating scales (e.g. percentage points); use single line texts with an open answer; in rank order questions and constant sum items, avoid having too many items to rank and items for point distributions respectively (see Chapter 24).

Some researchers approach survey companies to carry out their Internet survey. Using a professional company can address sampling matters; alternatively they may not perform very well, hence caution, 'vetting' and checking the company's previous experience are important. Denscombe (2014, p. 17) offers useful guidance for researchers who are considering using such services, addressing, for example: contracts and costs; security, privacy and sharing of data (particularly personal information) and trustworthiness; limits to the size of the survey; how the researcher accesses the data and in what format, and for how long the data can be accessed; how to prevent multiple submissions; password protection; design features (e.g. sample formats and templates); tracking (e.g. logged data on respondents and their contact details).

As the Internet becomes more popular for surveys, software resources for conducting these are becoming increasingly attractive and easy to use. Whilst there are plentiful advantages and considerations in Internet surveys, we also counsel researchers to be aware of the risks, ethics and considerations involved, as with all forms of research. Internet surveys have many advantages over their paper, telephone and face-to-face interview counterparts; they also bring their own concerns which we have addressed in this chapter.

18.8 Technological advances

In an era of rapid technological change it is invidious to be too prescriptive or narrow with regard to the technology for, design and conduct of, and access to, Internet surveys. Smartphones and mobile phone optimization, increasing user-friendliness, improved compatibility and integration between devices (and the Internet of Things), the ability of the same survey to be delivered in multiple formats to different devices, a huge and rapid increase in the range and types of mobile devices for accessing the Internet almost anywhere in the world and at any time, real-time communication, location software, immense strides in presentational and response software, increasing speed and connectivity, cloud computing, storage facilities for massive amounts of data, apps for everything and new social networking sites appearing almost daily, are all accumulating and advancing so quickly that even to name them here risks becoming instantly out of date. Simply keeping pace with developments is a full-time occupation. Researchers are advised to consult journals on digital technologies for social science research in order to keep up with the field.

Note

1 See, for example: Coomber (1997); Watt (1997); Dillman *et al*. (1998b, 2014); Dillman and Bowker (2000); Aldridge and Levine (2001); Roztocki and Lahri (2002); Archer (2003); Fox *et al*. (2003); Deutskens *et al*. (2005); Evans and Mathur (2005); Glover and Bush (2005); Joinson and Reips (2007); Fowler (2009); Bennett and Nair (2010); Farrell and Peterson (2010); Harlow (2010); Mora (2010, 2011a); Minnaar and Heystek (2013); Akbulut (2015); Diaz de Rada and Dominguez (2015).

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Case studies

CHAPTER 19

Case studies are important sources of research data, either on their own or to supplement other kinds of data, and constitute an approach to research in their own right. This chapter sets out key areas for attention in case studies:

- what is a case study?
- types of case study
- advantages and disadvantages of case study
- generalization in case study
- reliability and validity in case studies
- planning a case study
- case study design and methodology
- sampling in case studies
- data in case studies
- writing up a case study
- what makes a good case study researcher?

We provide researchers with an overview of key issues in the planning, conduct and reporting of a case study.

19.1 What is a case study?

It could be argued that any research in social science is a case. Case study might include experiment, action research, survey, naturalistic research, participatory research, historical research etc., and case study research uses multiple methods for data collection and analysis. In other words, it operates as many other types of research. Indeed Hamilton's and Corbett-Whittier's (2013) frequently cited *Case Study in Education Research* in many places reads like a general introductory volume on research methods and writing up research.

So a key question is 'what distinguishes a case study from other forms of research?' A case study has many definitions, indeed has been termed a 'contested terrain' (Yazan, 2015). For example, it has been defined as a specific instance that is frequently designed to illustrate a more general principle (Nisbet and Watt, 1984, p. 72). It is 'the study of an instance in action' (Adelman *et al*., 1980), 'the study of the particularity and complexity of a single case, coming to understand its activity within important circumstances' (Stake, 1995, p. xi). It is the 'detailed examination of a small sample' (Tight, 2010, p. 337) and an in-depth investigation of a specific, real-life 'project, policy, institution, program or system' from multiple perspectives in order to catch its 'complexity and uniqueness' (Simons, 2009, p. 21).

Whilst Creswell (1994, p. 12) defines the case study as a single instance of a bounded system, for example a child, a clique, a class, a school, a community, others would not hold to such a tight definition. For example, Yin (2009, p. 18) argues that the boundary line between the phenomenon and its context is blurred, as a case study is a study of a case in a context and it is important to set the case within its context (and rich descriptions and details are often a feature of a case study). Indeed Chong and Graham (2013) argue for a 'Russian doll' approach to understanding what a case study is, i.e. a nested approach where to understand a micro-level case involves understanding and including the meso- and macro-contextual levels in which it is nested (p. 24). A case study can sometimes be tightly bounded and other times less so; as Verschuren (2003, p. 123) argues, it is ambiguous.

Arriving at a single definition of case study is elusive and unnecessary. Is it, for example, a method, a process, a methodology, a research design, an outcome, a research strategy, a focus (Verschuren, 2003; Stake, 2005; Tight, 2010; Thomas, 2011; Yazan, 2015)? Ragin (1992) contrasts a case study approach to a 'variable' approach (p. 5), placing the case rather than specific variables at the heart of the study. In our comments below we attempt to address the several definitions of case study.

Equally taxing is defining what constitutes a 'case': whilst some authors define it as a bounded unit, this offers little purchase on our understanding, as it still does not define what constitutes a unit and what constitutes a boundary. Robson (2002, pp. 181–2) suggests that case study can include: an individual case study; a set of individual case studies; a social group study; studies of organizations and institutions; studies of events, roles and relationships. Punch (2005) notes that a case may be an individual, a group, an organization,

a community or a nation (p. 144) and Pring (2015) notes that a unit might be 'a person, institution or collection of institutions', for example, a School Board (p. 55). Indeed Tight (2010), quoting Punch (2005, p. 144), reports commentaries which argue that the unit of analysis (the 'case') in case studies is so unclear that 'almost anything can serve as a case', such that 'case study as a form of social research is not a particularly meaningful term' (Tight, 2010, p. 337) and can be replaced by terms such as 'small sample, in-depth study' (p. 38). We challenge this, arguing that researchers must make clear what their unit of analysis is, what is the level of their analysis, what constitutes the 'case', and what are their boundaries in case study research. Pring (2015) notes that, the larger the embrace of the unit, i.e. the wider the boundaries of the unit, the more complex becomes the task of unravelling, identifying and commenting on the interactions between all elements and levels of the unit.

A case study provides a unique example of real people in real situations, enabling readers to understand ideas more clearly than simply by presenting them with abstract theories or principles. Indeed a case study can enable readers to understand how ideas and abstract principles can fit together (Yin, 2009, pp. 72–3). Case studies can penetrate situations in ways that are not always susceptible to numerical analysis.

Case studies accept that there are many variables operating in a single case, and, hence, to catch the implications of these variables usually requires more than one tool for data collection and many sources of evidence. Case studies can blend numerical and qualitative data, and they are a prototypical instance of mixed methods research (see Chapter 2); they can explain, describe, illustrate and enlighten (Yin, 2009, pp. 19–20).

Verschuren (2003, p. 124), like many writers, argues that a distinguishing feature of case study research is 'holism' rather than 'reductionism'. Whilst for Yin (2009) 'holism' refers to conducting the research at the single unit of analysis chosen, which may be an individual, a group, an organization etc., for Verschuren the term 'holism' is ambiguous, as it may not necessarily mean looking at a *whole* subject, person, group or organization but only at the relevant areas of interest, taken together.

Case studies can establish cause and effect ('how' and 'why'); indeed one of their strengths is that they observe effects in real contexts, recognizing that context is a powerful determinant of both causes and effects, and that in-depth understanding is required to do justice to the case. As Nisbet and Watt (1984, p. 78) remark, the whole is more than the sum of its parts. Sturman (1999, p. 103) argues that a distinguishing feature of case studies is that human systems have a wholeness or integrity to them rather than being a loose connection of traits, and necessitate in-depth investigation. Contexts are unique and dynamic, hence case studies investigate and report the real-life, complex, dynamic and unfolding interactions of events, human relationships and other factors in a unique instance. Hitchcock and Hughes (1995, p. 316) suggest that case studies are distinguished less by the methodologies that they employ than by the subjects/objects of their inquiry, and there is frequently a resonance between case studies and interpretive methodologies. They further suggest (p. 322) that the case study approach is particularly valuable when the researcher has little control over events, i.e. behaviours cannot be manipulated or controlled (though some case studies, e.g. of therapies, may involve high levels of control).

Hitchcock and Hughes (1995, p. 317) consider that a case study has several hallmarks:

- it is concerned with a rich and vivid description of events relevant to the case;
- it provides a chronological narrative of events relevant to the case;
- it blends description with analysis of events;
- it focuses on individual actors or groups of actors, and seeks to understand their perceptions of events;
- it highlights specific events that are relevant to the case;
- the researcher is integrally involved in the case, and the case study may be linked to the personality of the researcher (cf. Verschuren, 2003, p. 133);
- an attempt is made to portray the richness of the case in writing up the report.

Similarly, Denscombe (2014) comments that case studies are characterized by: in-depth study of one setting; a focus on processes, interactions and relationships; holism; a concern for the particular; multiple methods of data collection; and focus on natural settings (pp. 54–7). Hamilton and Corbett-Whittier (2013, p. 11) add to this that case study: has its own approach to research (its own genre); has many contextual levels, from local to national; catches the complexity of a situation or context; may collect data on a single occasion or over time; often requires the researcher to spend time 'within the world of those being researched' (p. 11); and involves more than one perspective.

Case studies are set in temporal, geographical, organizational, institutional and other contexts that enable boundaries to be drawn around the case. They can be defined with reference to characteristics defined

by individuals and groups involved, and can be defined by participants' roles and functions in the case. Hitchcock and Hughes (1995) note that case studies:

- have temporal characteristics which help to define their nature;
- have geographical parameters allowing for their definition;
- have boundaries which allow for definition;
- can be defined by an individual in a particular context, at a point in time;
- can be defined by the characteristics of the group;
- can be defined by role or function;
- can be shaped by organizational or institutional arrangements.

Bassey (1999) comments that case studies in education can be conducted in order to inform decision making by policy makers, practitioners and theorists. They investigate 'interesting aspects of an educational activity, programme, or institution, or system … mainly in its natural context and within an ethic of respect for persons' such that plausible, trustworthy explanations and interpretations can be offered after collecting sufficient data in exploring the 'significant features of the case' (p. 58).

Case studies have the advantage over historical studies of including direct observation and interviews with participants (Yin, 2009, p. 11). They strive to portray 'what it is like' to be in a particular situation, to catch the close-up reality, rich detail and 'thick description' (Geertz, 1973) of participants' lived experiences of, thoughts about, and feelings for, a situation. They involve looking at a case or phenomenon in its real-life context, usually employing many types of data (Robson, 2002, p. 178). They are descriptive and detailed, with a narrow focus, and combine subjective and objective data (Dyer, 1995, pp. 48–9). It is important in case studies for events and situations to be allowed to speak for themselves, rather than to be heavily interpreted, evaluated or judged by the researcher.

This is not to say that case studies are unsystematic or merely illustrative; case study data are gathered systematically and rigorously. Indeed Nisbet and Watt (1984, p. 91) specifically counsel case study researchers to avoid:

- *journalism* (picking out more striking features of the case, thereby distorting the full account in order to emphasize these more sensational aspects);
- *selective reporting* (selecting only that evidence which will support a particular conclusion, thereby misrepresenting the whole case);
- *an anecdotal style* (degenerating into an endless series of low-level banal and tedious illustrations that take over from in-depth, rigorous analysis), i.e. the tendency of some case studies to overemphasize detail to the detriment of seeing the whole picture;
- *pomposity* (striving to derive or generate profound theories from low-level data, or by wrapping up accounts in high-sounding verbiage);
- *blandness* (unquestioningly accepting only the respondents' views, or only including those aspects of the case study on which people agree rather than areas on which they might disagree).

A key feature of case study is its rejection of a single reality; rather, there are multiple, multivalent realities operating in a situation, and the researcher's view and interpretation is only one of many. Indeed the researcher has a duty to address reflexivity and to address or report others', for example, participants' views on the case in question.

19.2 Types of case study

There are several types of case study. These can be determined by their *purposes*, for example, Denscombe (2014, p. 57): 'discovery-led' purposes which utilize description, exploration, comparison and explanation, and 'theory-led' purposes which utilize illustration and experiment. Yin (2009) identifies three types in terms of *outcomes*:

i *exploratory* (as a pilot to other studies or research questions). Exploratory case studies can be used to generate hypotheses that are tested in larger-scale surveys, experiments or other forms of research, for example, observational. However, Adelman *et al*. (1980) caution against using case studies solely as preliminaries to other studies, for example, as pre-experimental or pre-survey; rather, they argue, case studies exist in their own right as a significant and legitimate research method;
ii *descriptive* (providing narrative accounts);
iii *explanatory* (testing theories).

Yin's classification accords with Merriam (1998), who identifies three types:

i *descriptive* (e.g. narrative accounts);
ii *interpretative* (developing conceptual categories inductively in order to examine initial assumptions);
iii *evaluative* (explaining and judging).

Merriam also categorizes four common domains or kinds of case study: ethnographic, historical, psychological

and sociological. Sturman (1999, p. 107), echoing Stenhouse (1985), identifies four kinds of case study: an ethnographic case study (single in-depth study); action research case study; evaluative case study; and educational case study. Stake (1994, 1995, 2005) identifies three main types of case study:

i *intrinsic* case studies (studies that are undertaken in order to understand the particular case in question);
ii *instrumental* case studies (examining a particular case in order to gain insight into an issue or a theory);
iii *multiple/collective* case studies (groups of individual studies that are undertaken to gain a fuller or more general picture).

Hamilton and Corbett-Whittier (2013, pp. 15–19) deliberately move beyond Stake's 'intrinsic' and 'instrumental' types to suggest:

- *reflexive case study*: which includes the personal reflections of the researcher as the case/practitioner in question (raising concerns about personal bias, the need for outside perspectives and different data streams, and ethical issues with regard to colleagues);
- *longitudinal case study*: to catch changes over time, the dynamics of evolving situations and a sense of the history of an event or events, and to work with the same or different cohorts of participants (requiring sustained commitment and dedication to hard work, flexibility in design and data collection, and acceptance of changes over time);
- *cumulative case study*: case study or studies which provide a cumulative body of data about a topic, phenomenon or situation;
- *collective case study*: working separately and sometimes asynchronously to gather data about a particular phenomenon, situation or topic (e.g. a curriculum innovation);
- *collaborative case study*: working with others within and across institutions, to gather multiple perspectives and contexts.

Because case studies provide fine-grain detail, they can also be used to complement other, more coarsely grained – often large-scale – kinds of research. Case study materials in this sense can provide powerful human-scale data on macro-political decision making, fusing theory and practice.

Thomas (2011) and Thomas and Myers (2015) set out a clear and useful elements of case studies, which feature:

i the *subject*: whom and what to focus on, derived from local knowledge, a key case or an outlier case, for example, a deviant case; and the *object*: 'what is this a case of?' (Thomas, 2011, p. 515), what it is that has to be explained and in which the researcher is interested, the analytical issue that the researcher is exploring, i.e. the *explanandum*;
ii the *purpose* of the research: (the type of case study, e.g. intrinsic, instrumental, evaluative, exploratory);
iii the *approach* to be used: the kind of study, for example, theory-testing, theory-building, illustrative, descriptive, the *explanans* (the explanation or type of explanation or study to be used);
iv the *process* to be adopted: (a) a single case study (which may be retrospective, a 'snapshot') (p. 517), a diachronic study (a longitudinal study of change over time); (b) multiple cases (which might focus on: 'nested' cases, e.g. classes within a single school in which the school is the main case; 'parallel' cases which use several cases running simultaneously and independently; and 'sequential' cases with cases running consecutively, with one case affecting the subsequent case).

19.3 Advantages and disadvantages of case study

Case studies have several claimed strengths and weaknesses which have been identified for many years. Some of these are summarized in Box 19.1 (Adelman *et al*., 1980) and Box 19.2 (Nisbet and Watt, 1984).

Wellington (2015) adds to their strengths that they are illustrative and illuminating, accessible and easily disseminated, holding the reader's attention and being vivid accounts which are 'strong on reality' (p. 174). On the other hand, he notes that they are not replicable, may not be representative, typical or generalizable (p. 174). Denscombe (2014) notes the difficulties in choosing, knowing and setting boundaries to the case study, gaining access to case study settings and ensuring, where relevant, that case studies move beyond description to analysis and evaluation (p. 64).

Shaughnessy *et al*. (2003, pp. 290–9) suggest that case studies often lack a high degree of control, and treatments are rarely controlled systematically and have little control over extraneous variables. This, they argue, renders it difficult to make inferences and to draw cause and effect conclusions from case studies, and there is potential for bias in some case studies as the researcher might be both participant and observer and may overstate or understate the case (verification bias). Case studies, they argue, may be impressionistic, and self-reporting may be biased (by the participant or

BOX 19.1 POSSIBLE ADVANTAGES OF CASE STUDY

Case studies have a number of advantages that make them attractive to educational evaluators or researchers. Thus:

1. Case study data, paradoxically, are 'strong in reality' but difficult to organize. In contrast, other research data are often 'weak in reality' but susceptible to ready organization. This strength in reality is because case studies are down-to-earth and attention-holding, in harmony with the reader's own experience, and thus provide a 'natural' basis for generalization.
2. Case studies allow generalizations either about an instance or from an instance to a class. Their peculiar strength lies in their attention to the subtlety and complexity of the case in its own right.
3. Case studies recognize the complexity and 'embeddedness' of social truths. By carefully attending to social situations, case studies can represent something of the discrepancies or conflicts between the viewpoints held by participants. The best case studies are capable of offering support to alternative interpretations.
4. Case studies, considered as products, may form an archive of descriptive material sufficiently rich, varied and complex to admit subsequent reinterpretation.
5. Case studies are 'a step to action'. They begin in a world of action and contribute to it. Their insights may be directly interpreted and put to use: for staff or individual self-development; for within-institutional feedback; for formative evaluation; and in educational policy making.
6. Case studies present research or evaluation data in a more publicly accessible form than other kinds of research report, although this virtue is to some extent bought at the expense of their length. The language and the form of the presentation is hopefully less esoteric and less dependent on specialized interpretation than conventional research reports. The case study is capable of serving multiple audiences. It reduces the dependence of the reader upon unstated implicit assumptions and makes the research process itself accessible. Case studies, therefore, may contribute towards the 'democratization' of decision making (and knowledge itself). At their best, they allow readers to judge the implications of a study for themselves.

Source: Adapted from Adelman *et al*. (1980)

BOX 19.2 NISBET AND WATT'S (1984) STRENGTHS AND WEAKNESSES OF CASE STUDY

Strengths

1. The results are more easily understood by a wide audience (including non-academics) as they are frequently written in everyday, non-professional language.
2. They are immediately intelligible; they speak for themselves.
3. They catch unique features that may otherwise be lost in larger-scale data (e.g. surveys); these unique features might hold the key to understanding the situation.
4. They are strong on reality.
5. They provide insights into other, similar situations and cases, thereby assisting interpretation of other similar cases.
6. They can be undertaken by a single researcher without needing a full research team.
7. They can embrace and build in unanticipated events and uncontrolled variables.

Weaknesses

1. The results may not be generalizable except where other readers/researchers see their application.
2. They are not easily open to cross-checking, hence they may be selective, biased, personal and subjective.
3. They are prone to problems of observer bias, despite attempts made to address reflexivity.

Source: Adapted from Nisbet and Watt (1984)

the observer). Further, they argue that bias may be a problem if the case study relies on an individual's (selective) memory.

Dyer (1995, pp. 50–2) remarks that, reading a case study, one has to be aware that a process of selection has already taken place, and only the author knows what has been selected in or out, and on what criteria, and indeed the participants themselves may not know what selection has taken place. Indeed he observes (pp. 48–9) that case studies combine knowledge and inference, and it is often difficult to separate these; the researcher has to be clear about which of these feature in the case study data.

Case studies frequently follow the interpretive tradition of research – seeing the situation through the eyes of participants – rather than the quantitative paradigm, though this need not always be the case. Its sympathy to the interpretive paradigm has rendered case study an object of criticism. For example, Smith (1991, p. 375) argues that not only is the case study method the logically weakest method of knowing, but that studying individual cases, careers and communities is passé, and that attention should be focused on patterns and laws in historical research.

This is prejudice and ideology, perhaps, but it signifies the problem of respectability and legitimacy that case study had to conquer. Like other research methods, case study has to demonstrate reliability and validity. This can be difficult, for given the uniqueness of situations and multiple realities and perspectives, case studies may be, by definition, inconsistent with other case studies or unable to demonstrate this positivist view of reliability. Even though case studies are not obliged to demonstrate this form of reliability, nevertheless there are important questions to be faced in undertaking them, for example (Adelman *et al*., 1980; Nisbet and Watt, 1984; Hitchcock and Hughes, 1995):

- What exactly is a case?
- How are cases identified and selected?
- What kind of case study is this (what is its purpose)?
- What is reliable evidence?
- What is objective evidence?
- What is an appropriate selection to include from the wealth of generated data?
- What is a fair and accurate account?
- Under what circumstances is it fair to take an exceptional case (or a critical event – see the discussion of observation in Chapter 26)?
- What kind of sampling is most appropriate?
- To what extent is triangulation required and how will this be addressed?
- Triangulation seeks to determine a single, fixed point; what if the case study is characterized by many changing points, perspective and interpretations?
- What is the nature of the validation process in case studies?
- How will the balance be struck between uniqueness and generalization?
- What is the most appropriate form of writing up and reporting the case study?
- What ethical issues are exposed in undertaking a case study?

19.4 Generalization in case study

It is often heard that case studies, being idiographic, have limited generalizability (e.g. Yin, 2009, p. 15). Of course, the same could be said of single experiments (p. 15) and other kinds of research. Ruddin (2006) questions whether generalizability is an appropriate requirement of case study at all (p. 798), as it connotes positivism in what is not a positivistic type of research.

However, just as the generalizability of single experiments can be extended by replication and multiple experiments, so, too, case studies can be part of a growing pool of data, with multiple case studies contributing to greater generalizability. However, more pertinent is the claim by Robson (2002, p. 183) and Yin (2009, p. 15) that case studies opt for 'analytic' rather than 'statistical' generalization.

In statistical generalization the researcher seeks to move from a sample to a population, based on sampling strategies, frequencies, statistical significance and effect size. However, in analytic generalization, the concern is not so much for a representative sample (indeed the strength of the case study approach is that the case only represents itself) so much as its ability to contribute to the expansion and generalization of theory (Yin, 2009, p. 15) which can help researchers to understand other similar cases, phenomena or situations, i.e. there is a *logical* rather than *statistical* connection between the case and the wider theory. Yin (p. 43) makes the point that to assume that generalization is only from sample to population/universe is simply irrelevant, inappropriate and inapplicable to case studies. Rather, he argues (pp. 38–9) that case studies can help to generalize to a broader *theory* which can be tested in one or more empirical cases (akin, in this respect, to a single experiment or quasi-experiment) and can be shown *not* to support rival, even if plausible, theories.

Generalization requires extrapolation, and the case study researcher, whilst not necessarily being able to extrapolate on the basis of typicality or representativeness, nevertheless can extrapolate to relevant theory

(Macpherson *et al*., 2000, p. 52) and to the 'broader class' (Ruddin, 2006, p. 799) and to the testing or falsification of theory (it only takes one counter-example to disprove a theory: sighting one black swan negates the theory that all swans are white).

Case studies can make theoretical statements, but, like other forms of research and human sciences, these must be supported by the evidence presented. This requires the nature of generalization in case study to be clarified. Generalization can take various forms, for example:

- from the single instance to the class of instances that it represents (e.g. a single-sex selective school might act as a case study to catch significant features of other single-sex selective schools);
- from features of the single case to a multiplicity of classes with the same features;
- from the single features of part of the case to the whole of that case;
- from a single case to a theoretical extension or theoretical generalization.

A robust defence of generalization from case studies is made by Verschuren (2003, p. 136). First, he argues that statistical generalization is made on the basis of the homogeneity (or variability) of the population and the sample, together with the level of certainty required in the sample (see Chapter 12). So, for example, if the population is highly standardized and invariant (he uses the example of a factory that makes the same, uniform, standardized machines) the sample used for quality control could well be small, whereas in a very variable population (with many variables with a range of values in each) the sample size would have to be large. He then turns to the number of case studies which might be required for generalizability to be secure, and he argues that, in fact, a very small number of case studies could be used, each of which embraces the range of variables in question, thereby reducing the number of overall cases required; this is because 'complex issues in general have a much lower variability than separate variables' (p. 137), i.e. the researcher can generalize from a small number of case studies that represent the complex issues in general (cf. Pring, 2015, p. 57). The argument is clear: case studies include many variables; multivariable phenomena are often characterized by homogeneity rather than high variability; therefore if the researcher can identify case studies that catch the range of variability then external validity – generalizability – can be demonstrated.

A further case for generalizability from case studies (Watts, 2007; Thomas, 2010; Simons, 2015) argues that, in fact, there are universals present in each case (the case study carries 'exemplary knowledge' of a wider phenomenon) (Thomas, 2010, p. 576), even though case study is the study of the singular and the unique (Simons, 2015, p. 175). Here the narrative style which often characterizes case studies enables the reader to connect their own experiences to those reported in the case study. Simons writes that we can learn from a specific, single and singular case where it promotes 'generalized understanding' (p. 174) and offers something of 'universal significance' (p. 181). Despite our manifest differences, we gain universal understanding vicariously from single case studies, just as with poems, novels and short stories (p. 175), and apply them to our own situation (p. 177). She notes that this is nothing new; humans have been 'generalizing from the particular' (p. 184) from time immemorial. Similarly, Thomas (2010) and Thomas and Myers (2015) note that case study does not need to conform to the scientific notion of generalizability but, rather, to the contribution that it makes to the understanding and practical wisdom (phronesis) of the researcher and reader. This is echoed by Pring (2015), who notes that, rather than there being generalizability in the scientific sense, case studies can 'alert one to similar possibilities in other situations. They, as it were, "ring bells" ' (p. 56).

19.5 Reliability and validity in case studies

Whilst case studies may not have the external checks and balances found in other forms of research, nevertheless they still abide by canons of validity and reliability, for example:

- *construct validity* (through employing accepted definitions and constructions of concepts and terms; operationalizing the research and its measures/criteria acceptably);
- *internal validity* (through ensuring agreements between different parts of the data, matching patterns of results, ensuring that findings and interpretations derive from the data transparently, that causal explanations are supported by the evidence (alone), and that rival explanations and inferences have been weighed and found to be less acceptable that the explanation or inference made, again based on evidence);
- *external validity* (clarifying the contexts, theory and domains to which generalization can be made);
- *concurrent validity* (using multiple sources and kinds of evidence to address research questions and to yield *convergent validity*, e.g. triangulation of

data, investigators, perspectives, methodologies, instruments, time, location, contexts);
- *ecological validity* (fidelity to the special features of the context in which the study is located);
- *reliability* (replicability and internal consistency);
- *avoidance of bias* (e.g. the case study simply being an embodiment or fulfilment of the researcher's initial prejudices or suspicions, with selective data being gathered or data being used selectively (Yin, 2009, p. 72), or with the researcher's bias being inevitable if the researcher is a participant observer whose personality may affect the research process (Verschuren, 2003, p. 122). This can be addressed by reflexivity, respondent checks or checks by external reviewers of the data, inferences and conclusions drawn).

Of note here is Yin's (2009, pp. 41, 122–4) call for a 'chain of evidence' to be provided, such that an external researcher can track through every step of the case study from its inception to its research questions, design, data sources, instrumentation, data (evidence and the circumstances in which they were collected, e.g. time, place and functional interconnections of people, places etc.) and conclusions. It is important to note the time and place in which case study data are collected, as many actions and events are context-specific and part of a 'thick description', as this enables replication research to be planned (Macpherson *et al*., 2000, p. 56).

19.6 Planning a case study

In planning a case study there are several issues that researchers can consider:

- The particular circumstances of the case, including: (a) the possible disruption to individual participants that participation might entail; (b) negotiating access to people; (c) negotiating ownership of the data; (d) negotiating release of the data.
- The conduct of the study including: (a) the use of primary and secondary sources; (b) the opportunities to check data; (c) triangulation (including peer examination of the findings, respondent validation and reflexivity); (d) data-collection methods (in the interpretive paradigm case studies tend to use certain data-collection methods, e.g. semi-structured and open interviews, observation, narrative accounts and documents, diaries, maybe also tests, rather than other methods, e.g. surveys, experiments. Nisbet and Watt (1984) suggest that, in conducting interviews, it may be wiser to interview senior people later rather than earlier to make maximum use of discussion time with them, the interviewer having been put into the picture fully before the interview); (e) data analysis and interpretation, and, where appropriate, theory generation; (f) the writing of the report, with conclusions separated from the evidence, with essential evidence included in the main text, and balancing illustration with analysis and generalization.
- The consequences of the research (for all parties). This might include the anonymizing of the research in order to protect participants, though such anonymization might suggest that a primary goal of case study is generalization rather than the portrayal of a unique case, i.e. it might go against a central feature of case study. Anonymizing reports might render them anodyne, and the distortion that may be involved in such anonymization to render cases unrecognizable might be too high a price to pay for going public. Is it realistic and/or desirable not to identify the case and participants? Researchers must ensure that due concern has been given to ethical matters; this continues right through the case study period, from planning to conducting to reporting (see Chapter 7).

Thomas and Myers (2015) suggest that, in planning a case study, researchers must consider whether the case study is singular or multiple. They need to focus on intuition, understanding, theorization and analysis and, using thick descriptions, connect analysis with explanations.

Nisbet and Watt (1984, p. 78) suggest three main stages in undertaking a case study. Because case studies catch the dynamics of unfolding situations it is advisable to commence with a very wide field of focus, an open phase, without selectivity or pre-judgement. Thereafter 'progressive focusing' enables a narrower field of focus to be established, identifying key foci for subsequent study and data collection. At the third stage a draft interpretation is prepared which needs to be checked with respondents before appearing in the final form. The authors (p. 79) advise against generating hypotheses too early in a case study; rather, they suggest, it is important to gather data openly.

Respondent validation can be particularly useful as respondents might suggest a better way of expressing the issue or may wish to add or qualify points. There is a risk in respondent validation, however, that they may disagree with an interpretation. Nisbet and Watt (1984, p. 81) indicate the need to have negotiated rights to veto. They also recommend that researchers consider: (a) promises that respondents can see those sections of the report that refer to them (subject to controls for

confidentiality, e.g. of others in the case study); (b) take full account of suggestions and responses made by respondents and, where possible, to modify the account; (c) in the case of disagreement between researchers and respondents, promise to publish respondents' comments and criticisms alongside the researchers' report.

Sturman (1997) places on a set of continua the nature of data collection, data types and data-analysis techniques in case study research. These are presented in summary form in Table 19.1.

At one pole are unstructured, typically qualitative data, whilst at the other are structured, typically quantitative data. Researchers using case study approaches will need to decide which methods of data collection, which type of data and techniques of analysis to employ, all on the basis of fitness for purpose.

In planning a case study Thomas (2011) makes an important distinction between the *subject* and *object* of a case study. The subject is the example, the focus (e.g. an education system, a school, a group of students), whereas the object is that which has to be explained – the *explanandum* – for instance, the structures, management effectiveness and levels of achievement respectively. Selecting the subject – the focus – of the case is a matter of sampling, and we discuss sampling below (e.g. critical cases, extreme cases, typical cases). Taking into account Thomas's distinction of *subject* and *object*, in planning the study the researcher will need to consider:

- the most appropriate *subject* (focus) of the study in order to address the purpose, for example, a group of students, a particular child, a group of teachers, a curriculum innovation etc. (e.g. derived from local knowledge, key cases or outlier cases);
- the *object* of the study: what it is that has to be explained in which the researcher is interested, the analytical issue that the researcher is exploring, the *explanandum*;
- the *purpose* of the study (e.g. intrinsic, instrumental, evaluative, exploratory), addressing fitness for purpose;
- the *approach* to be used, for example, theory-testing, theory-building, illustrative/descriptive;
- the *process* to be used, for example, single (retrospective, snapshot, diachronic) or multiple (nested, parallel, sequential);
- the *sample* (e.g. a critical case, an extreme case, a typical case, a representative case).

For example, the researcher might be interested in why upper secondary school male students outperform female students in science subjects. This is the *explanandum*, that which is to be explained, i.e. the *object* of the researcher: 'what is this a case of' (Thomas, 2011, p. 515). The researcher decides that the most suitable focus here is Form 5 and Form 6 male and female students' results; this is the *subject* of the research. The researcher decides the most effective kind of case study (e.g. an exploratory case study) and approach to be used (e.g. theory-building). Then the researcher decides on the sampling strategy, and she adopts a 'typical case' sampling, involving those males and females who do and do not decide to follow sciences in post-school study or employment, those who are following discipline-specific science (physics, chemistry, biology) and General Science in Forms 5 and 6, the careers guidance teachers and science teachers in the school, and the parents of the students in question.

TABLE 19.1 CONTINUA OF DATA COLLECTION, TYPES AND ANALYSIS IN CASE STUDY RESEARCH

	Data Collection	
Unstructured (field notes)	⟷ (interviews – open to closed)	Structured (survey, census data)
	Data Types	
Narrative (field notes)	⟷ (coded qualitative data and non-parametric statistics)	Numeric (ratio scale data)
	Data Analysis	
Journalistic (impressionistic)	⟷ (content analysis)	Statistical (inferential statistics)

Source: Adapted from Sturman (1997)

Hamilton and Corbett-Whittier (2013, pp. 51–62) identify six 'key decisions' in approaching the planning of a case study:

i 'self-reflection' (where you actually are) (p. 53);
ii 'research questions' (where you wish to go) (p. 55);
iii 'defending your methodological approach' (p. 57);
iv 'strategic approaches' ('who will do what, when and with whom') (p. 59);
v 'getting organized' ('what will go where, when') (p. 60);
vi 'presenting the findings' (p. 61).

Stake (2005) argues that the qualitative case study should include (pp. 459–60):

- setting the boundaries of the case and conceptualizing the object of study;
- selecting appropriate phenomena, issues or themes for study, which might be framed in the research questions;
- seeking patterns in the data in order to develop the issues of focus;
- triangulation of key observations in order to support interpretations;
- identifying alternative interpretations for further study;
- developing generalizations or assertions from the case.

19.7 Case study design and methodology

Yin (2009, pp. 46ff.) identifies four main case study designs:

i The *single-case design* can focus on a critical case, an extreme case, a unique case, a representative or typical case, a revelatory case (an opportunity to research a case heretofore unresearched, e.g. Whyte's *Street Corner Society*, see Chapter 15), a longitudinal case.
ii The *embedded single-case design*, in which more than one 'unit of analysis' is incorporated into the design, for example, a case study of a whole school might also use sub-units of classes, teachers, students, parents, and each of these might require different data-collection instruments, for example, a survey questionnaire, interviews, observations etc.
iii The *multiple-case design*, for example, comparative case studies within an overall piece of research, or replication case studies. Campbell (1975, p. 180) suggests that having two case studies, for comparative purposes, is more than worth having double the amount of data on a single case study! For example, educationists may want to see the effects of a new innovation, let us say in mathematics teaching, in three circumstances (conditions): one where teachers are given in-house staff development for the new mathematics, one where they attend externally provided courses on the new mathematics, and another where the teachers receive both kinds of staff development; here the case studies might compare the effects in the schools concerned (cf. Yin, 2009, pp. 54–5).
iv The *embedded multiple-case design*, in which different sub-units may be involved in each of the different cases and a range of instruments (e.g. a survey questionnaire, interviews, observations, archival records etc.) might be used for each sub-unit, and each is kept separate to each case.

A single case may be part of a multiple case-study design, and, by contrast, a particular data-collection instrument (e.g. a survey) may be part of a cross-site case study. In considering multiple case studies, it is important to decide how many are required; typically, the more subtle is the issue under investigation, the more cases are required (Yin, 2009, p. 58) in order to be able to rule out rival explanations. Yin also notes that a single-case design can overlook the possible benefits of multiples cases, for example, replication, thereby avoiding the criticism of being a unique, single case in which the researcher is 'putting all the eggs in one basket', which may be risky: an 'all-or-nothing' risk.

A key issue in case study research is the selection of information. Though it is frequently useful to record typical, representative occurrences, the researcher need not always adhere to criteria of representativeness. For example, it may be that infrequent, unrepresentative but critical incidents or events occur that are crucial to the understanding of the case. A subject might only demonstrate a particular behaviour once, but it is so important as not to be ruled out simply because it occurred once; sometimes a single event might occur which sheds a hugely important insight into a person or situation (see the discussion of critical incidents in Chapter 33); it can be a key to understanding a situation (Flanagan, 1949; Tripp, 1993).

For example, it may be that a psychological case study might happen upon a single instance of child abuse earlier in a person's life, but the effects of this are so profound as to constitute a turning point in understanding that adult. A child might suddenly pass a single comment that indicates complete frustration with

or complete fear of a teacher, yet it is too important to overlook. Case studies, in not having to seek frequencies of occurrences, can replace quantity with quality and intensity, separating the significant few from the insignificant many instances of behaviour. Significance rather than frequency is a hallmark of case studies, offering the researcher an insight into the real dynamics of situations and people.

In designing a case study, Yin (2009, p. 27) indicates five components to address:

- the case study's questions (it was suggested earlier that case study is particularly powerful in answering the 'how' and 'why' type of questions, and Yin (p. 29) argues that the more specific are the questions that the case study should answer, the stronger is the likelihood of the case study staying on track and within limits);
- the case study's propositions (if there are any) (e.g. a hypothesis to be tested);
- the case study's 'unit(s) of analysis' (this relates to the key issue in case study, which is defining what constitutes the case, e.g. an individual, a group, a community, an organization, a programme, a piece of innovation, a decision and its ramifications, an industry, an economy etc.). What constitutes the case should be clear from the research questions being asked (p. 30), as these should specify the unit of analysis. Yin (p. 32) suggests that the unit of analysis should be concrete (a real-life phenomenon) rather than abstract (e.g. an argument or topic). Identifying the unit of analysis can be used to identify the tricky question of what constitutes a case;
- the logic that links the data gathered to the propositions set out in the case study (i.e. how the data will be analysed, e.g. by looking for patterns, explanations, analysis of events as they unravel over time, cross-site and cross-case analysis) (p. 34);
- the 'criteria for interpreting the findings' from the case study (which includes a clear indication of how the interpretation given is better than rival explanations of the data).

Yin (p. 35) also adds that theory generation should be included in the research design phase of the case study, as this assists in focusing the case study; such theories might be of the behaviour of individuals, groups, organizations, communities, societies, i.e. there are several levels of theory.

Unlike the experimenter who manipulates variables to determine their causal significance or the survey researchers who ask standardized questions of large, representative samples of individuals, the case study researcher typically observes the characteristics of an individual unit – a child, a clique, a class, a group, a school or a community. The purpose of such observation is to probe deeply and to analyse intensively the multi-stranded phenomena that constitute the life of the unit, possibly with a view to generalizing to the wider population to which that unit belongs.

Observation in case study

Case studies are methodologically eclectic (i.e. embedded within them may be more than one kind of research such as ethnography, experiment, action research, survey, illuminative research, observational research, documentary research); they can use a range of methods of data collection, data types (quantitative and qualitative) and ways of analysing data (statistically and through qualitative tools), and they can be short term or long term. In short, case study is a hybrid (cf. Verschuren, 2003, p. 125). That said, at the heart of many case studies lies observation.

Case studies vary in their degree of structure, for example, 'natural' (e.g. ethnographies) to artificial (e.g. a counselling situation, the Stanford Prison Experiment and the Milgram studies of obedience (see Chapter 30)); structured (e.g. structured non-participant observations) to unstructured (e.g. ethnographic observation); interventionist (e.g. a case study of an individual undergoing therapy) to non-interventionist (e.g. a child study).

There are two principal types of observation: participant observation and non-participant observation. In the former, observers engage in the very activities they set out to observe. Often their 'cover' is so complete that as far as the other participants are concerned, they are simply one of the group. In the case of Patrick, for example, born and bred in Glasgow, his researcher role remained hidden from the members of the Glasgow gang in whose activities he participated for four months (Patrick, 1973). Such complete anonymity is not always possible, however. Thus in Parker's study of downtown Liverpool adolescents, it was generally known that the researcher was waiting to take up a post at the university. In the meantime, 'knocking around' during the day with the lads and frequenting their pub at night rapidly established that he was 'OK'. The researcher was, in his own terms, 'a drinker, a hanger-arounder' who could be relied on to keep quiet on illegal matters (Parker, 1974).

Cover is not necessarily a prerequisite of participant observation. In a study of a small group of working-class boys during their last two years at school and their first months in employment, Willis (1977) attended all the different subject classes at school – 'not as a

teacher, but as a member of the class' – and worked alongside each boy in industry for a short period.

Non-participant observers, on the other hand, stand aloof from the group activities they are investigating and eschew group membership. For example, the non-participant observer role is where the researcher sits at the back of a classroom writing notes or coding up the verbal exchanges between teacher and students onto structured observational categories.

Bailey (1994, p. 247) explains that it is hard for a researcher who wishes to undertake covert research not to act as a participant in a natural setting, as, if the researcher does not appear to be participating, then why is he/she there? Hence, in many natural settings the researchers are participants. This is in contrast to laboratory or artificial settings, in which non-participant observation (e.g. through video recording) may take place.

The unstructured, ethnographic account of teachers' work is a typical method of observation in the natural surroundings of a setting, for example, a school in which the study is conducted. Similarly, structured observations may be a common approach in a more artificial setting, for example, a counsellor's office.

The natural scientist, Schutz (1962) points out, explores a field that means nothing to the molecules, atoms and electrons therein. By contrast, the subject matter of the world in which the educational researcher is interested is composed of people and is essentially meaningful to them. That world is subjectively structured, possessing particular meanings for its inhabitants. The task of the educational investigator is often to explain the means by which an orderly social world is established and maintained in terms of its shared meanings. How do participant observation techniques assist the researcher in this task? Bailey (1994, pp. 243–4) identifies some inherent advantages in the participant observation approach:

1. Observation studies are superior to experiments and surveys when data are being collected on non-verbal behaviour.
2. In observation studies, investigators are able to discern ongoing behaviour as it occurs and are able to make appropriate notes about its salient features.
3. Because case study observations take place over an extended period of time, researchers can develop more intimate and informal relationships with those they are observing, generally in more natural environments than those in which experiments and surveys are conducted.
4. Case study observations in natural settings are less reactive than other types of data-gathering methods. For example, in laboratory-based experiments and surveys that depend upon verbal responses to structured questions, bias can be introduced in the very data that researchers are attempting to study.

Further, direct observation is faithful to the real-life, *in situ* and holistic nature of a case study (Verschuren, 2003, p. 131).

19.8 Sampling in case studies

Sampling has a dual meaning here: the participants in the case study, or the kind of case study to be adopted. With regard to the latter is the decision about purposive sampling: whether to choose a typical case, a representative case, a critical case, an extreme case, a deviant case, an outlier, intensity sampling, maximum variation sampling (e.g. for multiple case studies), homogeneous sampling, reputational case sampling, revelatory case sampling, theoretical sampling, opportunistic sampling etc. We review these in Chapter 12, and we advise readers to go to this chapter. At issue here is the need for the selection of the case to be fit for purpose, relevant to the topic or issue in hand, to include the significant features of the subject and object of the research, to be a suitable instance of the phenomenon under investigation, to be suitably bounded and to be capable of maintaining a holistic view of the case as well as its particular contributing elements.

Having decided the most suitable *kind* of case, the researcher then turns to the most appropriate sampling of *people* or *issues*. Here, again, the researcher can utilize typical case sampling, and case studies often use non-probability, purposive samples (see Chapter 12). Again, the researcher must select the sample for the case study in terms of fitness for purpose.

Often the case study and its participants are chosen as being 'typical cases', critical cases or extreme cases. Robson (2002, pp. 181–2) notes the distinction between a critical case study and an extreme or unique case. In a critical case study, the case in question might possess all, or most, of the characteristics or features that one is investigating, more fully or distinctly than under 'normal' circumstances, for example, a case study of student disruptive behaviour might go to a *very* disruptive class, with students who are very seriously disturbed or challenging, rather than going into a class where the level of disruption is not so marked.

By contrast, Robson argues (2002, p. 182) that the extreme and the unique case can provide a valuable 'test bed'. Extremes include, he argues, the situation in which 'if it can work here it will work anywhere', or choosing an ideal set of circumstances in which to try

out a new approach or project, maybe to gain a fuller insight into how it operates before taking it to a wider audience (e.g. the research and development model).

19.9 Data in case studies

We mentioned earlier that case studies are eclectic in the types of data that are used. Indeed many case studies will rely on mixed methods and a variety of data. Whilst observation and participant observation are often pre-eminent in case studies, they are by no means the only sources of data. For example, Yin (2009, p. 101) identifies:

- *documents* (p. 103), for example, letters, emails, memoranda, agendas, minutes, reports, records, diaries, notes, other studies, newspaper articles, website uploads, etc.;
- *archival records* (p. 105), for example, public records, organizational records and reports, personal (maybe medical or behavioural) and personnel data stored in an organization (with due care to privacy legislation), charts and maps;
- *interviews* (p. 106): in-depth, focused, and formal survey interviews (see Chapter 25);
- *direct observation* (p. 109), i.e. non-participant observation of the natural setting and the target individual(s), groups *in situ*, artefacts, rooms, decor, layout;
- *participant observation* (p. 111), in which the researcher takes on a role in the situation or context featured in the case study;
- *physical artefacts* (p. 113), for example, pictures, furniture, decorations, photographs, ornaments.

Here the multiple sources of evidence can provide convergent and concurrent validity on a case, and they demand of the researcher an ability to handle and synthesize many kinds of data simultaneously. This, in turn, advocates the compilation of a case study database of evidence (Yin, 2009, p. 118) that comprises two main kinds of collection: the actual data gathered, recorded and organized by entry, and the researcher's ongoing analysis, report, comments and narrative on the data.

The diverse data provide the evidence needed for the researcher to draw conclusions, the evidential 'chain of evidence' that gives credibility, reliability and validity to the case study (Yin, 2009, p. 122). When writing the report, the researcher must make direct reference to the actual evidence that supports the point being made, and we turn to the writing of the case study report below.

The researcher can use several computer-assisted software tools (e.g. NVivo) to process the data ready for analysis (see Chapter 32). These can group, retrieve, organize and search single and multiple data sets, and return these ready for analysis and presentation in such forms as (Miles and Huberman, 1984):

- matrices and arrays of data;
- patterns, themes and configurations;
- narratives;
- data displays;
- flowcharts;
- within-site and cross-site analyses;
- cause and effect diagrams and chains (e.g. where an effect becomes a subsequent cause);
- networks of relationships and causes or linked events (i.e. rather than linear models of cause and effect);
- chronologies and causal sequences;
- time series and critical events;
- key issues and subordinate issues;
- explanations;
- tabulations;
- grounded theory.

Yin (2009, p. 143) makes the point that, in analysing data, the researcher has to go back through the data several times to ensure that all the data fit the interpretations given or conclusions drawn, i.e. without unexplained anomalies or contradictions (the constant comparison method), that all the data are accounted for (p. 160), that rival interpretations are considered (p. 160) and that the significant features of the case are highlighted (p. 161). It may be that there are several perspectives and interpretations of the data, as case studies deal in multiple realities rather than a single right answer.

The recording of observations is a frequent source of concern to inexperienced case study researchers. Whilst field notes in ethnographic research are typically copious, how much should be recorded, and in what form? What does one do with the mass of recorded data? We offer several suggestions here with regard to field notes:

- record the notes as quickly as possible during or after observation, since the quantity of information forgotten is very slight over a short period of time but accelerates quickly over time;
- discipline yourself to write notes quickly and reconcile yourself to the fact that recording field notes can take as long as time spent in actual observation, and transcribing interviews can take four or five times

longer than the actual interview, so use transcription sparingly;
- recording and dictating rather than writing may be possible but writing has the advantage of stimulating thought;
- entering field notes onto a secure computer file is preferable to handwriting, as it is easy to store, recover, read, process and manipulate data;
- field notes should be sufficiently full and vivid to make sense after time has passed (e.g. after a month or months).

Field notes are often part of unstructured observation studies. Such notes, confessed Wolcott (1973), helped him fight the acute boredom that he sometimes felt when observing the interminable meetings that were the daily lot of the school principal. Occasionally, however, a series of events would occur so quickly that Wolcott had time only to make cursory notes which he supplemented later with fuller accounts. One useful tip from this experienced ethnographer is worth noting: never resume your observations until the notes from the preceding observation are complete. There is nothing to be gained merely by your presence as an observer. Until your observations and impressions from one visit are a matter of record, there is little point in returning to the classroom or school and reducing the impact of one set of events by superimposing another and more recent set. Indeed, when to record one's data is but one of a number of practical challenges identified by Walker (1980), which are listed in Box 19.3.

19.10 Writing up a case study

Writing up a case study abides by the twin criteria of 'fitness for purpose' and 'fitness for audience'. Robson (2002, pp. 512–13) and Yin (2009, pp. 176–9) suggests six forms of organizing the writing-up of a case study:

1. In the *suspense structure* the author presents the main findings (e.g. an executive summary) in the opening part of the report and then devotes the remainder of the report to providing the evidence, analysis, explanations, justifications (e.g. for what is selected in or out, what conclusions are drawn, what alternative explanations are rejected) and argument that lead to the overall picture or conclusion.
2. In the *narrative report* a prose account is provided, interspersed with relevant figures, tables, emergent issues, analysis and conclusion.
3. In the *comparative structure* the same case is examined through two or more lenses (e.g. explanatory, descriptive, theoretical) in order either to provide a rich, all-round account of the case, or to enable the reader to have sufficient information from which to judge which of the explanations, descriptions or theories best fit(s) the data.
4. In the *chronological structure* a simple sequence or chronology is used as the organizational principle, enabling cause and effect to be addressed and possessing the strength of an ongoing story. The chronology can be sectionalized as appropriate (e.g. key events or key time frames), and can intersperse

BOX 19.3 THE CASE STUDY AND PROBLEMS OF SELECTION

Among the issues confronting the researcher at the outset of his case study are the problems of selection. The following questions indicate some of the obstacles in this respect:

1. How do you get from the initial idea to the working design (from the idea to a specification, to usable data)?
2. What do you lose in the process?
3. What unwanted concerns do you take on board as a result?
4. How do you find a site which provides the best location for the design?
5. How do you locate, identify and approach key informants?
6. How they see you creates a context within which you see them. How can you handle such social complexities?
7. How do you record evidence? When? How much?
8. How do you file and categorize it?
9. How much time do you give to thinking and reflecting about what you are doing?
10. At what points do you show your subject what you are doing?
11. At what points do you give them control over who sees what?
12. Who sees the reports first?

Source: Adapted from Walker (1980)

commentaries on, interpretations of, explanations for, and summaries of emerging issues as events unfold (e.g. akin to 'memoing' in ethnographic research). The chronology becomes an organizing principle, but different kinds of contents are included at each stage of the chronological sequence.

5 In the *theory-generating structure*, the structure follows a set of theoretical constructs or a case that is being made. Here, Robson suggests, each succeeding section of the case study contributes to, or constitutes, an element of a developing 'theoretical formulation', providing a link in the chain of argument, leading eventually to the overall theoretical formulation.

6 In the *unsequenced structures* the sequence, for example, chronological, issue-based, event-based, theory-based, is unimportant. Robson suggests that this approach renders it difficult for the reader to know which areas are important or unimportant, or whether there are any omissions. It risks the caprice of the writer.

Some case studies are of a single situation – a single child, a single social group, a single class, a single school. Here any of the above six approaches may be appropriate. Some case studies require an unfolding of events, others operate under a 'snapshot' approach (e.g. of several schools, or classes, or groups at a particular point in time). In the former it may be important to preserve the chronology, whereas in the latter such a chronology may be irrelevant. Some case studies are divided into two main parts (e.g. Willis, 1977): the data reporting and then the analysis/interpretation/explanation.

A case study report should consider rival explanations of the findings and indicate how the explanation adopted is better than its rivals. Such rival explanations might include, for example (Yin, 2009, pp. 133–5):

- the role of chance/coincidence;
- experimenter effects or situation effects (reactivity);
- researcher bias;
- other influences on the case;
- covariance or the influence of another variable, i.e. a cause other than the intervention or situation reported explains the effects;
- alternative explanations of what the data show;
- the *process* of the intervention, rather than its *contents*, explain the outcome;
- a different theory can explain the findings more fully and fittingly;
- the intervention was part of a much bigger intervention that was already taking place at the time of the case study, so is subsumed by that bigger intervention;
- observed changes might have happened anyway, without the intervention from the case study.

Yin (2009, pp. 185–9) suggests that an 'exemplary' case study must be 'significant', 'complete', take into consideration 'alternative perspectives', be careful to include 'sufficient evidence' and be 'engaging'. These precepts, surely, can provide a useful guide for researchers.

19.11 What makes a good case study researcher?

A case study requires in-depth data, a researcher's ability to gather data that address fitness for purpose, and skills in probing beneath the surface of phenomena. These requirements imply that the researcher must be an effective questioner, listener, prober, able to make informed inferences (to 'read between the lines'; Yin, 2009, p. 70) and adaptable to changing and emerging situations. Given that a case study uses a range of methods for data collection (e.g. observation (participant to non-participant), accounts, interviews, artefacts, documents, archival records, survey), and that it may use different methodologies within it (e.g. action research, experiment, ethnography), the effective case study researcher must be versed in each of these, know how to draw on them at the most appropriate moment, be able to keep a clear sense of direction in the data collection, so that the case study is kept on track and not side-tracked, and have a clear grasp of the issues for which the case study is being conducted (and keep to these). Clarity of focus, issues and direction are important here.

Further, the effective case study researcher will need to possess the ability to collate and synthesize data from different sources, to make inferences and interpretations based on evidence, to know how to test inferences and conclusions (and how to test them against rival explanations) and know how to report multiple perspectives.

The case study researcher is often privy to confidential or sensitive material. Hence he/she must be clear on: the ethics of the research; his/her own stance in respect of disclosing private or sensitive data; how to protect people at risk or vulnerable groups; how to address matters of justified covert research; whether to report people anonymously or to identify them; how to address non-traceability and non-identifiability of participants; non-attributability of particular comments to individuals; and how to incorporate specific, important features into a cross-site analysis.

It is important for the case study researcher to have the subject knowledge and research expertise required to conduct the case study, to be highly prepared, to have a sense of realism about the situation being researched (as case study is a 'real-life' exercise), to be an excellent communicator (which may require training) and to have the appropriate personality characteristics that will enable access, empathy, rapport and trust to be built up with a diversity of participants. Not every researcher has all of these, yet each is vitally important.

Finally, case study researchers, like other educational researchers, are concerned with providing factual information, explanations and theories rather than, for example, the promotion of their own value judgements (Foster *et al*., 2000). Their value judgements do not have any privileged position, taking into account, of course, that intellectual authority and expertise may be important. Of course, factual information may be value-relevant, but that is not the same as making value judgements (pp. 22–3).

19.12 Conclusion

Macpherson *et al*. (2000, pp. 57–8) set out several principles to guide the practice of case study research. With regard to *purpose*, they suggest a collaborative approach between participants and researcher in order to address *contextuality*. With regard to *place*, they suggest *sensitivity* to the place (akin to ecological validity). With regard to both *purpose* and *process*, they suggest *authenticity* (fitness for purpose), *applicability* (thinking large but starting small) and *growth* (ensuring development and social transformation). With regard to *product*, they suggest *communicability* of the findings through networking (which they also apply to *purpose* and *process*).

Case study has had a mixed press. Flyvberg (2006), Yin (2009) and Ulriksen and Dadalauri (2016), for example, note that it has been regarded as a weaker sibling to other methods because of its putative loose structure, limited generalizability, biased case selection which derives from knowledge of the dependent variable, informality and indiscipline, limited empirical legitimacy, subjectivity and subjective conclusions, but Pring (2015, p. 56) argues that this is to falsely assume that there exists a single reality rather than multiple realities.

Thomas (2010) notes that in case study, as in science more widely, which uses induction, rather than expecting permanent universality or generalizability (which is a misplaced hope), 'exemplary knowledge' is more suited to the phronesis of case study and to multiple interpretations and horizons of researchers and readers of case study. Further, Morrison (2009) and Ulriksen and Dadalauri (2016) note that case studies have considerable potential for providing causal explanations (Ulriksen and Dadalauri (2016) develop this in terms of 'process tracing').

The authors referenced in this chapter have powerfully and roundly refuted the putative weaknesses of case study and have accorded it a place alongside and equal to other kinds of research in social science and educational research. We hold to this latter position: case study has a unique and distinctive contribution to make to educational research. Whether to use case study is driven by fitness for purpose.

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Experiments

CHAPTER 20

This chapter discusses key issues in experiments in education, indicating how they might address causality as a main target of much educational research. The chapter includes:

- randomized controlled trials
- designs in educational experiments
- true experimental designs
- quasi-experimental designs
- single-case ABAB design
- procedures in conducting experimental research
- threats to internal and external validity in experiments
- the timing of the pre-test and the post-test
- the design experiment
- Internet-based experiments
- *ex post facto* research

The intention here is to introduce different forms of experiment, to ensure that researchers are aware of key issues to be addressed in their planning and conduct, and what might or might not legitimately be inferred from their results.

20.1 Introduction

Experiments, particularly – indeed sometimes exclusively – randomized controlled trials (RCTs), in educational research seem unstoppable, rapidly achieving hegemonic status (Pearce and Raman, 2014). From being a matter of under-representation in educational research in the early part of the century, their allure now seems irresistible to governments and researchers alike (cf. National Research Council, 2002), and the Campbell Collaboration (the social science equivalent of the Cochrane Collaboration in medicine) provides powerful evidence of this, including the provision of research syntheses and meta-analyses. The literature is replete with examples of experiments, and a cursory Internet search will return thousands of examples.[1]

Experiments make several claims (cf. Denscombe, 2014): scientific credibility, repeatability, precision and causality. The great claim of experimental methods, particularly RCTs, is that they demonstrate causality, i.e. that an outcome has been caused by a specific intervention. The issue of *causality* and, hence, predictability has exercised the minds of researchers (Morrison, 2009), and one response has been in the operation of *control* of variables and settings, and it finds its apotheosis in experimental design. If rival causes or explanations can be eliminated from a study then clear causality can be established; the model can *explain* outcomes causally. The National Research Council (2002), Torgerson and Torgerson (2008), Torgerson (2009) and Morrison (2009) note that the experimental approach concerns itself with causality; this is contestable, as we make clear in Chapter 6.

The essential feature of experimental research is that investigators deliberately control and manipulate the conditions which determine the events in which they are interested, introduce an intervention and measure the difference that it makes. An experiment involves making a change in the value of one variable – the independent variable – and observing the effect of that change on another variable – the dependent variable. Experimental research can be *confirmatory*, seeking to support or not to support a null hypothesis, or *exploratory*, discovering the effects of certain variables. In an experiment the post-test measures the dependent variable, and the independent variables are isolated and controlled carefully.

20.2 Randomized controlled trials

Randomized controlled trials (RCTs), a 'true' experiment (discussed below), have considerable prominence in education; hence we devote much discussion to them in this chapter. Experiments inform policy and practice in education, and, as Torgerson (2009) notes, if they are sufficiently large, can take account of different characteristics of students, the nature and implementation of an intervention, and differences in outcome (p. 314).

The US has the 'What Works Clearinghouse' and the Institute of Education Sciences (IES) which report RCTs. The What Works Clearinghouse enables educationists to interrogate the data of RCTs in education by topic, student characteristics and units of randomization

(individual and cluster). International organizations focus on RCTs (e.g. Bouguen and Gurgand (2012) report national RCTs in Europe), as do educational researchers in the evidence-based movement (e.g. Torgerson and Torgerson, 2001, 2003a, 2013; Moore *et al*., 2003; Gorard and Torgerson, 2006; Hutchison and Styles, 2010; Goldacre, 2013; Hassey, 2015).

In the UK, the Educational Endowment Foundation was established in 2011 (Torgerson and Torgerson, 2013), initiating fifty-nine RCTs involving 2,300 schools; the Behavioural Insights Team opened in 2012; and in 2013 the UK's Department for Education announced two major RCTs on (a) schools' attainment in mathematics and science and (b) child protection. Haynes *et al*. (2012), in a publication issuing from the Cabinet Office of the UK government, declared that '[r]andomised controlled trials (RCTs) are the best way of determining whether a policy is working' (p. 4). This echoes statements elsewhere that RCTs provide 'the best scientific evidence' on policies such as educational technology, class size and school vouchers (Angrist, 2003, p. 1).

In order to increase the explanatory power of RCTs in education – why certain effects are found – they are often accompanied by ethnographic data ('process evaluations').

Key elements of a randomized controlled trial

Imagine that we have been transported to a laboratory to investigate the properties of a new wonder-fertilizer that farmers could use on their cereal crops (and agriculture was an early user of RCTs), let us say wheat (Morrison, 1993, pp. 44–5, based on Fisher, 1966). The scientist would randomly take from a bag of wheat seed a number of seeds and then randomly split them into two equal parts. One part would be grown under normal existing conditions: controlled and measured amounts of soil, warmth, water and light, with other factors excluded. This would be called the control group. The other part would be grown under the same conditions: the same controlled and measured amounts of soil, warmth, water and light as the control group, *and*, additionally, the new wonder-fertilizer. Then, four months later, the two groups are examined and their growth measured. The control group has grown half a metre and each ear of wheat is in place but the seeds are small. The experimental group, by contrast, has grown half a metre as well but has significantly more seeds on each ear, and the seeds are larger, fuller and more robust.

The scientist concludes that, because both groups came into contact with nothing other than measured amounts of soil, warmth, water and light, then it could not have been anything else but the new wonder-fertilizer that caused the experimental group to flourish so well. The key factors in the experiment were:

- the random selection of the seeds from a population of seeds;
- the random allocation of the randomly selected sample of wheat into two matched groups (the control and the experimental group), involving the initial measurement of the size of the wheat to ensure that it was the same for both groups (i.e. the pre-test);
- the identification and isolation of key variables (soil, warmth, water and light);
- the control of the key variables (the same amounts to each group);
- the exclusion of any other variables;
- the giving of the special treatment (the intervention) to the experimental group (i.e. manipulating the independent variable) whilst holding every other variable constant for the two groups;
- ensuring that the two groups are entirely separate throughout the experiment (non-contamination);
- the final measurement of yield and growth to compare the control and experimental groups and to look at differences from the pre-test results (the post-test);
- the comparison of one group with another;
- the stage of generalization – that this new wonder-fertilizer improves yield and growth under a given set of conditions.

In educational research this translates into:

- random sampling of participants from a population;
- random allocation of the sample to control or experimental groups;
- pre-testing the control and experimental groups to ensure parity, i.e. that there are no statistically significant differences or large effect sizes between them;
- identification and isolation of key variables;
- control of the key variables;
- exclusion of any other variables;
- special treatment (the intervention) given to the experimental group (i.e. manipulating the independent variable) whilst holding every other variable constant for the two groups;
- ensuring that the two groups are entirely separate throughout the experiment (non-contamination);
- final measurement of outcomes to compare the control and experimental groups and to look at differences from the pre-test results (the post-test);

- comparison of one group with another, to see the effects of the intervention on the experimental groups and the dependent variable.

The RCT – the 'true' experiment – is represented diagrammatically in Figure 20.1.

So strong is this simple and elegant 'true' experimental design, that all the threats to internal validity identified in Chapter 14 are, according to Campbell and Stanley (1963), controlled in the pre-test–post-test control group design. The term 'control' has been used in two main senses so far: the random allocation of participants to a control or an experimental group and the isolation and control of variables. Whilst the former is self-evident, in the second the researcher isolates key independent variables and controls what happens to these, for example, so that the same amounts of these are given to both the control group and the experimental group, i.e. the control group and experimental groups are matched in their exposure to these independent variables. This involves giving an identical, measured amount of exposure of both groups to these (whether this can actually be achieved in practice is a moot point, but for the purpose of the discussion here we assume it can). By holding the independent variable constant (giving the same amount to both the control group and the experimental group), it is argued that any changes brought about in the experimental group must be attributable to the intervention, the other variables having been held constant (controlled).

The importance of randomization

Schneider *et al*. (2007, p. 13) suggest that Holland's (1986) 'fundamental problem of causal inference' (a person cannot be in both the control and the experimental group simultaneously) comes into being once one accepts that a causal effect is the difference between what would have happened to a person in an experiment if she had been in the experimental group (receiving the intervention) and if the same person had been in the control group. This 'fundamental problem' is addressed through randomization, and a key feature of an RCT is, as its name suggests, randomization:

> Randomization is a key, critical element of the 'true' experiment; random sampling and random allocation to either a control or experimental group is a key way of allowing for the very many additional uncontrolled and, hence, unmeasured, variables that may be part of the make-up of the groups in question.… It is an attempt to overcome the confounding effects of exogenous and endogenous variables: the *ceteris paribus* condition (all other things being equal); it assumes that the distribution of these extraneous variables is more or less even and perhaps of little significance. In short it strives to address Holland's (1986) 'fundamental problem of causal inference', which is that a person may not be in both a control group and an experimental group simultaneously.… [B]ecause random allocation takes into account both observed and unobserved factors, controls on unobserved factors, thereby, are unnecessary.… If students are randomly allocated to control and experimental group and are equivalent in all respects (by randomization) other than one group being exposed to the intervention and the other not being exposed to the intervention, then, it

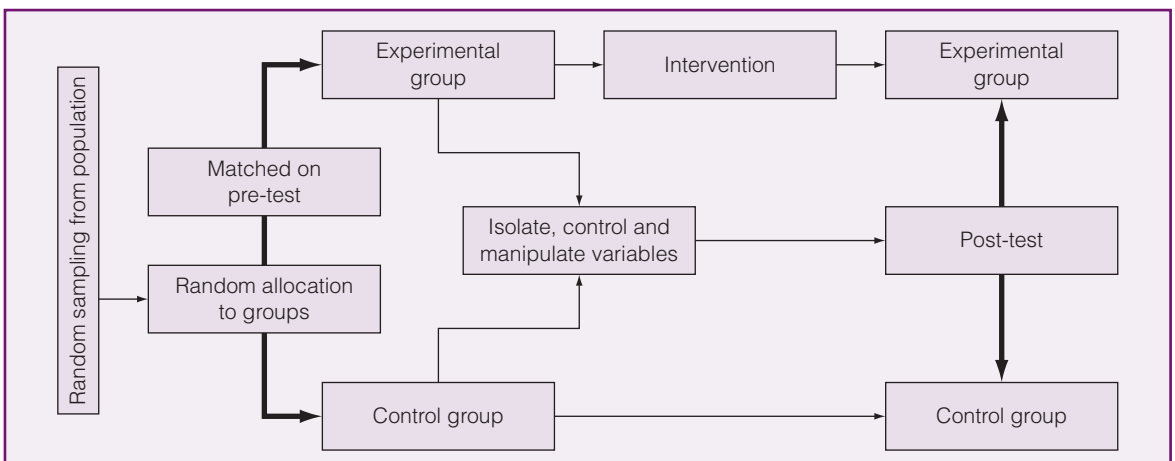


FIGURE 20.1 The 'true' experiment

> is argued, the researcher can attribute any different outcomes between the two groups to the effects of the intervention.
>
> (Morrison, 2009, pp. 143–4)

Kerlinger (1970) observes that, in theory, random assignment to experimental and control groups controls all possible independent variables. In practice, of course, it is only when enough subjects are included in the experiment that the principle of randomization has a chance to operate as a powerful control. However, the effects of randomization even with a small number of subjects is well illustrated in Box 20.1.

Randomization ensures the greater likelihood of equivalence, that is, the equal apportioning out between the experimental and control groups of any other factors or characteristics of the subjects which might conceivably affect the experimental variables in which the researcher is interested (cf. Torgerson and Torgerson, 2003a, 2003b, 2008). If the groups are equivalent, then any 'clouding' effects (other minor variables) should be present in both groups.

Randomization, Smith (1991, p. 215) explains, produces equivalence over a whole range of variables, whereas matching produces equivalence over only a few named variables. Randomization is a way of reducing the effects of allocation bias (Sullivan, 2011), ensuring that baseline features or characteristics, which may not be known to the researcher, are evenly distributed between the control and experimental groups.

Holland (1986, p. 947) suggests a statistical solution to his 'fundamental problem of causal inference' through randomization and the measurement of the *average* results (p. 948). The *average* score on the pre-test and post-test may be useful unless it masks important differences between subsets of the two samples, for example, students with a high IQ and students with a low IQ may perform very differently, but this would be lost in an average, in which case stratification into sub-samples can be adopted. We address problems of averages below.

Schneider *et al*. (2007, pp. 13–15) also make suggestions to address Holland's problem:

- Place the same person in the control group, followed by placing her in the experimental group (which assumes *temporal stability* (cf. Holland 1986, p. 948), i.e. the fact that there are two time periods must make no difference to the results, there being a constancy of response, regardless of time), assuming or demonstrating that the placement of the person in the first group does not affect the person for long enough to contaminate (affect) the person's response to being in the second group (cf. Holland 1986, p. 948) (see below: repeated measures designs).
- Assume that all the participants are identical in every respect (which may be possible in the physical sciences but questionably so in the human sciences, even in studies of twins (Holland, 1986, p. 947)).

Torgerson (2009) notes that, in educational research, randomization may occur at the class or school level rather than the individual person level, as the individuals in a class are not completely independent of each other, i.e. there may be a bias in only working within individuals in a single class or in a single school. Cluster sampling also reduces the risk of contamination (the experimental group influencing the control group and vice versa) which may occur if the trial is contained

BOX 20.1 THE EFFECTS OF RANDOMIZATION

Select twenty cards from a pack, ten red and ten black. Shuffle and deal into two ten-card piles. Now count the number of red cards and black cards in either pile and record the results. Repeat the whole sequence many times, recording the results each time.

You will soon convince yourself that the most likely distribution of reds and blacks in a pile is five in each: the next most likely, six red (or black) and four black (or red); and so on. You will be lucky (or unlucky for the purposes of the demonstration!) to achieve one pile of red and the other entirely of black cards. The probability of this happening is 1 in 92,378! On the other hand, the probability of obtaining a 'mix' of not more than six of one colour and four of the other is about 82 in 100.

If you now imagine the red cards to stand for the 'better' ten children and the black cards for the 'poorer' ten children in a class of twenty, you will conclude that the operation of the laws of chance alone will almost probably give you close equivalent 'mixes' of 'better' and 'poorer' children in the experimental and control groups.

Source: Adapted from Pilliner (1973)

within a single school. Cluster sampling means that the number of individuals in the sample increases significantly in order to ensure statistical power (see Chapter 39), as each class or school becomes just one cluster – one unit – which, in turn, comprises individuals (p. 316).

For example, Torgerson and Torgerson (2008, p. 100) suggest that a new curriculum is implemented at the whole school level, rather than an individual person level. Hence the unit of randomization is the school, so the researcher would have to randomly sample, from the population of schools, several schools to be the control group and several other schools to be the experimental group. This might present problems in finding sufficient schools, as it increases the sample size, each school counting as only one unit (Tymms (2012) reports the example of using 120 schools in one project that used cluster sampling). Torgerson (2009) suggests that it is preferable to use many small schools, each with a small number of students, rather than a smaller number of schools with large numbers of students in each. This echoes the comment of Lindquist (1940) that 'the *unit* of sampling in educational research' may be the class or the school, or indeed the community, rather than the student (p. 24).

Cluster sampling, however, reduces the chance of finding a difference between the control and experimental groups. It may 'dilute any intervention effects' (Torgerson and Torgerson, 2008, p. 100) and it may risk bias in choosing the individual people from a cluster. The authors note also that statistical treatment in cluster samples may be more sophisticated, as it may use multilevel modelling (though Gorard (2013, p. 107) argues against this). Further, cluster sampling runs the risk that, since the unit of analysis is the school and not the individual, as individuals come and go in any one school, it may be that the post-test is conducted on students who were not included in the pre-test. This problem can be attenuated by ensuring that random sampling of individuals takes place at the pre-test and post-test stages. For further analysis of cluster-level analysis we refer the reader to Torgerson and Torgerson (2008) and Bland (2010).

Full randomization (i.e. random *sampling*: selection from a total population) in much educational research may be impracticable, even impossible (Lindquist, 1940, pp. 24–5), but random *allocation* may be possible (e.g. within a school), and, as Lindquist notes, this may suffice for adherence to sampling theory.

Concerns about randomized controlled trials

Powerful advocacy of RCTs for planning and evaluation is provided by Boruch (1997), Torgerson and Torgerson (2008) and Goldacre (2013). Indeed Boruch argues (1997, p. 69) that the problem of poor experimental controls has led to highly questionable claims being made about the success of programmes.

RCTs in education have their protagonists and antagonists. On the one hand, RCTs claim to provide evidence of 'what works', which is preferable to introducing or using untested interventions in education. RCTs can meet a rigorous standard of evidence and can upset long-held, false myths about education, and can suggest probabilistic causation. Small-scale RCTs acting as pilots can also reduce risk.

On the other hand, RCTs have been criticized on many counts. For example, the irreducible complexity and multiplicity of purposes, contexts and changing dynamics of participants in a specific context (Brooks *et al*., 2014, p. 71), intended outcomes and contents of education frustrate the simplicity of RCTs. Concerns have also been raised about the questionable ethics of randomization (e.g. denying control groups access to potentially positive interventions). Further, randomization in educational RCTs might be difficult, and the solution may not necessarily be provided by cluster randomization (Torgerson, 2009).

The many challenges facing RCTs in education have also been well aired.[2] For example, classical experimental methods, abiding by the need for replicability and predictability, may not be particularly fruitful since, in complex phenomena, results are never clearly replicable or predictable: we never step into the same river twice. Further, in linear thinking, small causes bring small effects and large causes bring large effects, but, as in complexity and chaos theory, small causes can bring huge effects and huge causes may have little or no effect. Moreover, to atomize phenomena into measurable variables and then to focus only on certain ones of these is to miss synergy and the spirit of the whole. Measurement, however acute, may tell us little of value about a phenomenon; I can measure every physical variable of a person but the nature of the person, what makes that person who she or he is, eludes atomization and measurement. RCTs, in this sense, have to answer the sometimes discredited view of science as positivism.

The RCT, premised on notions of randomization, isolation and control of variables in order to establish causality, may be appropriate for a laboratory, though whether, in fact, a social situation either ever *could*

become the antiseptic, artificial world of the laboratory or *should become* such a world are empirical and moral questions respectively. Indeed, the discussion of the 'design experiment' later in this chapter notes that its early advocate (Brown, 1992) had moved away from laboratory experiments to naturalistic settings in order to catch the true interaction of myriad variables in the real world. Further, the ethical dilemmas of treating humans as manipulable, controllable and inanimate are considerable (see Chapter 7).

Whilst we address ethical concerns in Chapter 7, it is important here to note the common reservation that is voiced about the two-group experiment (e.g. Gorard, 2001b, p. 146), which questions how ethical it is to deny a control group access to a treatment or intervention in order to suit the researcher (to which the counter-argument is, as in medicine, that (a) the researcher does not know whether the intervention, e.g. the new drug, will work or whether it will bring harmful results, and indeed the purpose of the experiment is to discover this (Goldacre, 2013), and (b) if an intervention works, then it can be offered to the control group at a later date once the trial has finished).

Hage and Meeker (1988, p. 55) suggest that the experimental approach may be fundamentally flawed in assuming that a single cause produces an effect. Further, it may be that the setting effects are acting causally, rather than the intervention itself, i.e. where the results are largely a function of their context (see Maxwell, 2004), for instance in the Milgram studies of obedience and the Stanford Prison Experiment reported in Chapter 30 and Zimbardo (2007a, 2007b).

Morrison (2001, p. 69) argues that RCTs in education on their own operate from a restricted view of causality and predictability; understate the value of other data sources and types; display unrealistic reductionism, simplification and atomization of a complex whole; understate the importance of multiple perspective in judging 'what works'; fail to catch the dynamics of non-linear phenomena; are silent on the processes (and causal processes) that take place in experiments (the black box approach); and neglect the significance of context. In other words, undifferentiated RCTs alone cannot tell the whole story of efficacy, generalizability and effectiveness.

Whilst randomization, harking back to Fisher, is designed to overcome myriad within-group and between-group differences, focusing on average results of control and experimental groups, this might be all well and good for the agricultural model in *The Design of Experiments* (1966), but humans, for example, students in school, are infinitely more complex and less passive than seeds which are affected by soil, heat, light, weather, location and water. An educational intervention is not like putting a fertilizer onto a patch of soil; a fertilizer may have only one effect whereas education may have many, and, whereas fertilizers look for average effects, education concerns the benefits to individuals. Further, in education, one intervention may cause a multiplicity of outcomes and may vary according to the characteristics of the students. Indeed Tymms (1996) notes that the same treatment with the same class may produce different results.

The National Research Council (2002) notes that RCTs may be expensive, may lack generalizability and, anyway, 'cannot test complex causal hypotheses' (p. 125) (see also Cartwright and Hardie, 2012). On the other hand, Torgerson (2009) contends that RCTs are 'particularly well-suited to areas where there is considerable complexity in terms of causal pathways and mechanisms of action' (p. 314), as they override specific causal pathways, control out alternative explanations and concern themselves with an input and an outcome. Maxwell (2004) and Camburn *et al*. (2015) note that a significant shortcoming of an RCT is its failure to provide a causal basis for deciding *how* something works (p. 24) and how far it is generalizable.

One problem that has been identified with an RCT is the interaction effect of testing. Good (1963) explains that whereas the various threats to the validity of the experiments listed in Chapter 14 can be thought of as main effects, manifesting themselves in mean differences independently of the presence of other variables, interaction effects, as their name implies, are joint effects and may occur even when no main effects are present. For example, an interaction effect may occur as a result of the pre-test measure sensitizing the subjects to the experimental variable. Interaction effects can be controlled by adding to the pre-test–post-test control group design two more groups that do not experience the pre-test measures. The result is a four-group design, as suggested by Solomon (discussed below).

The RCT is the 'gold standard' of many educational researchers, as it purports to establish controllability, causality and generalizability (Coe *et al*., 2000; Curriculum, Evaluation and Management Centre, 2000). How far this is true is contested (Morrison, 2001). For example, complexity theory replaces simple causality with an emphasis on networks, linkages, holism, feedback, relationships and interactivity in context (Cohen and Stewart, 1995), emergence, dynamical systems, self-organization and an open system, rather than the closed world of the experimental laboratory (Morrison, 2012). Even if we could conduct an experiment, its applicability to ongoing, emerging, interactive, relational, changing, open situations, in practice, may be

limited (Morrison, 2001, 2012). It is misconceived to hold variables constant in a dynamical, evolving, fluid, open situation.

We also question whether the complexity of education lends itself to RCTs and we suggest that pragmatic, 'real-world' RCTs are more useful than laboratory-like trials, and in fact non-laboratory experiments are likely to be the only options in educational research. Indeed Campbell, a towering figure in experimental research in education, was an advocate of quasi-experiments and field experiments (Shadish *et al*., 2002; Pearce and Raman, 2014).

Some RCTs may have limited external validity (generalizability), and findings in one context may not work in another context (Cartwright and Hardie, 2012). Further, blind and double-blind experiments may not be feasible in education. An experiment may fail to catch, or may ignore, the complexity and significance of teacher–student interactions, and education comprises an ongoing, dynamic interplay of systems, contexts and people that may not be captured in a single RCT, i.e. RCTs over-simplify the 'real world' (cf. Hammersley, 2015b). Indeed Sullivan (2011) notes that contextual factors may trump the findings from RCTs, i.e. it may not be clear whether a result is due to the context or to the intervention, and this is particularly so if the intervention is 'fairly dilute' (p. 285).

Smith (2013) notes that it is difficult to operate RCTs in education because outcomes are not easy to predefine and, even if we could identify such outcomes in education, measuring them is challenging and surrogates and proxies may be problematic (a matter of construct validity, see Chapter 14).

What happens in the hermetically sealed world of the laboratory is unlike what happens in the 'real world' in which contamination and the Hawthorne effect may occur. An RCT might suggest whether something 'works', but not why or how, and these are what educationists need to know (Morrison, 2001; Pawson, 2013), for example, Camburn *et al*. (2015), studying an experiment on school principal training, found that the experiment 'did not illuminate why or how the program failed to influence principal practice' (p. 2).

There is a case for RCTs in educational research, but they must be rigorous, and whilst RCTs have their place, attention must also be given to: the 'real world' (and we explore field experiments and quasi-experiments later in this chapter); the whole person; context; differentiated sub-groups; differentiation by personal characteristics of participants; the amount, quality, strength, frequency, intensity and duration of an intervention (cf. Camburn *et al*., 2015, pp. 8–9) and the effects of differences in these on participants; recognizing that a person is a complex system which combines and connects very many elements whose interactions and outcomes change over time (with commensurate changes to interventions over time).

Further, whilst RCTs may have their place in educational research, this does not obviate the importance of, or preclude the use of, other research approaches (Marsden, 2007; Menter, 2013; Pring, 2015). Sheffield Hallam University (2016) echoes this:

> RCTs on their own provide limited detail on why an intervention has a positive (or negative) impact, or whether specific aspects of a complex intervention are more (or less) effective than others. Because of this, our RCT evaluation designs incorporate a process evaluation that mixes qualitative and quantitative research approaches.
>
> (Sheffield Hallam University, 2016, p. 1)

Pring (2015, p. 50) notes that Campbell himself, after whom the Campbell Collaboration is named, had reservations about the exclusion of qualitative research in the experimental approach. RCTs are only one source of evidence in educational research, and the argument has been advanced that they should be complemented by qualitative data of many different hues (e.g. Pring, 2015).

Whilst being able to identify whether an intervention 'works' under carefully controlled conditions, RCTs need to take account of 'real-world' settings, and improving RCTs involves sub-group identification and inclusion, in short, careful and detailed stratification and analysis of differential treatment effects. We are not against RCTs at all; the point is that if we wish to use RCTs in education they would benefit from greater rigour than often currently obtains.

The limits of averages

The measures used in RCTs focus on the average, overall results rather than outliers or important sub-sample differences (discussed below). Non-cognitive outcomes may be ignored, and focus is placed on whether a particular intervention brings its designed outcome, regardless of the cost (widely defined). Currently many RCTs in education are content with a single average measure, a single measure of effect size or statistical significance, overlooking intervention-response differences, within-group differences, between-group differences and sub-sample differences (which even factorial designs may not catch). This is an important feature here: if RCTs in education are to be conducted, then they need to be more sophisticated and, at the same time, sensitive to individual and group

differences, to context and to the need to move beyond singular measurements of outcomes such as averages. Average difference conceals within-individual and within-group variation, between-individual and between-group variation and interaction. Whilst stratification attenuates this, it increases the sample size (Chapter 12), for example, in each stratum, in order to retain statistical power.

On the other hand, Goldacre (2013), a protagonist for RCTs in education, remarks that in the world of education

> [e]very child is different, of course, and every patient is different too; but we are all similar enough that research can help find out which intervention will work best overall, and which strategies should be tried first, second or third, to help everyone achieve the best outcome.
>
> (Goldacre, 2013, p. 7)

Variable responses ('heterogeneity of treatment effects') are almost inevitable in heterogeneous individuals and their sub-groups. RCTs often overlook such heterogeneities, leading to claims for results being more broadly applicable than in reality they are. Indeed RCTs may overlook sub-groups, other conditions and interventions operating on the situation in question, and other students or contextual characteristics; this is a warning for educational research which too easily assumes that a single intervention will have a single effect; a blunt instrument with a blunt measure.

For RCTs in educational research, benefits come from the sophisticated sub-sampling and sub-group targeted treatments with varied outcomes, researched in suitably differentiated trials in the 'real world'. This is far from the relatively crude, undifferentiated input–output RCTs that appear in educational literature which typically report statistical significance and a single, overall effect size. Further, attention has to be given not only to the magnitude and nature of the effect but how this varies for individuals and sub-groups, arguing perhaps for factorial research designs in RCTs. Averages conceal such differences, and we argue here that RCTs have to be sensitive to variability in individuals and groups. Currently in education this is largely not the case.

Measures of central tendency in RCTs, typically by using averages, may be their strongest point or their Achilles heel, depending on the purpose of the research. Whereas RCTs may seek to establish the best intervention for the average student, and ignore outliers, education has a duty to attend to outliers, as students, be they average or outliers, may or may not benefit equally from the treatment. One should not assume population or response homogeneity, and it is all too easy to dismiss outliers, regardless of the levels at which they are defined: 1 per cent, 5 cent, 10 per cent, or whatever. In education, the more difficult, extreme and small in number is the sub-group, the more it risks being overlooked or even removed from the data analysis, yet it may be the outliers who benefit most from an intervention.

Many RCTs in education seem to take a largely undifferentiated approach to diagnosing who might and might not benefit from an intervention, and then proceed to a relatively crude RCT that is targeted at a relatively undifferentiated group. One lesson here is that educational RCTs may benefit from being differentiated to targeted groups, based on careful diagnosis; the other lesson is that interventions will need to take account of the whole person, not just a few variables.

Given the problems of using averages in RCTs, we suggest the benefits of supplementing the findings from RCTs with evidence from other methodologies and data, because excluding and including variables, i.e. focusing on a single variable of interest in an artificial setting, risks overlooking broader contexts and applicability. Similarly, focusing on one putative homogeneous group may misrepresent the nature of that group.

How do you know if the experiment has 'worked'?

It is often difficult to find an experiment (e.g. an RCT) in education that states in advance the level of success that it requires in order to be judged efficacious or effective and how it will address contingencies and responses to interventions. Many experiments have no clearly specified targets for effectiveness and efficacy, though some may indicate an *overall* effect size sought in order to ensure statistical power (cf. Ellis, 2010). Indeed it is difficult to find RCTs in education which state their prognoses, targeted improved benefit (for whom, how much and about what), predicted benefit or its lack (and for whom), predicted risk and its mitigation (though ethics committees are supposed to act here), and important details of the intervention.

Researchers want to know if their experiment has 'worked'. How can they be assured of this? There are several answers to this question. One approach is to use null hypothesis significance testing; another is to consider effect size; another is to address the statistical power of a test; another is to adopt the subtraction approach; another is to consider rival explanations of the findings; another is more complex, to recognize that unequivocal measures may not tell researchers all that they wish to know, and that they may wish to know the

contingencies and conditions under which their experiment did and did not work: the contingency approach (Pawson, 2013). Further, since experiments may have differing outcomes for different participants, this injects an ethical dimension into the experimental outcomes. We consider all these points below.

Null hypothesis significance testing

Null hypothesis significance testing (NHST), as we discuss in Chapter 39, strives to determine whether results found, for example, whether an intervention makes a difference, and is or is not by chance. It does not and cannot tell the researcher how much difference an intervention makes, and most researchers want to know this: the magnitude of an effect. In Chapter 39 we note the limits of NHST: it is silent on what many researchers and users of research want or need to know, i.e. how much of an effect an intervention has and on whom (which groups and sub-groups), under what conditions and contingencies, with how much 'treatment' (e.g. quantity, quality, intensity, strength, frequency, duration) and at what cost. Indeed we raise concerns about the assumptions on which NHST is built, for example, the assumption of the null hypothesis (Chapter 39).

Fisher's (1966) comment that randomization, linked to the importance of averages and intended to overcome a range of individual differences, 'will suffice to guarantee the validity of the test of significance, by which the result of the experiment is to be judged' (p. 21) is questionable, as significance testing has limited value, telling the researcher only about the likelihood of the result occurring by chance (and indeed this is questionable, see Chapter 39) rather than which students might or might not benefit, and by how much; one cannot read off from a general result or a significance test what will be the result for an individual.

Effect size

Effect size (e.g. Cohen's *d*) is a widely used measure of difference. Effect size is usually measured in standard deviation units, with different measures used for different numbers of groups (e.g. a two-group design; a design with more than two groups). Here researchers should specify in advance of the research what effect size they require in order to judge whether their experiment has 'worked', for example, whether they will be content with a low or medium effect size, or whether they really need a large effect size to warrant their judgement of success. We address effect size in Chapter 39.

Statistical power

Whether an experiment has 'worked' depends on the statistical power of the test, its ability to detect an effect if there is one, avoiding a Type I error – a false positive – and a Type II error – a false negative. In other words, statistical power suggests how much confidence we can place in the results. Researchers should specify what statistical power they wish, as this affects sample size, and this must be set before the research testing takes place. Too often researchers do not indicate the statistical power of the test, and often their sample size is so small that the statistical power is weak, and the results could be simply by chance. We address statistical power in Chapter 39.

The subtraction approach

In the subtraction approach the putative causal effect of an intervention is calculated thus:

Step 1: Subtract the pre-test score of the experimental group from the post-test score of the experimental group to yield score (1).
Step 2: Subtract the pre-test score of the control group from the post-test score of the control group to yield score (2).
Step 3: Subtract score (2) from score (1).

If the result is negative then the causal effect is negative. Though this approach is straightforward, it is difficult to interpret the results, as the criterion for judging success has to be made clear and has to be judged with reference to the scale being used. For example, if I follow the three steps outlined above and I find a difference of, say, 10 points, is this large or small? The answer to this depends on the scale being used: if the scale runs from 0 to 20 then the difference of 10 points is proportionally large; if the scale runs from 0 to 100 then the difference of 10 points is proportionally much smaller. The researcher will need to decide the appropriate level (e.g. proportion of difference) for judging whether the experiment has 'worked'. Further, though this approach is intended to show the average causal effect, a figure on its own does not determine causality; rather it is the design of the experiment itself that may affect any inferences of probabilistic causality.

Considering rival explanations

Like statistical power, this approach is designed to enable the researcher to know how much confidence can be placed in the results obtained. Here the researcher has to consider alternative, rival explanations for the findings, and then defend the claim that these rival explanations are not as persuasive as the

interpretation proffered, for example, that the intervention has not only caused the observed finding, rather than other factors, but has also caused the magnitude of the observed finding. This depends in part on the warrants being brought forward to support the conclusion reached (see Chapter 11), and in part on the power of the evidence brought forward.

The contingency approach

In this approach researchers want more than a simple metric of how much difference an intervention has made, whether this was by chance and how much confidence they can place in the result. Rather, the researcher wishes to know under what circumstances and conditions (contingencies) it works or does not work, for whom and in what terms and under what criteria. Pearce and Raman (2014), commenting on the relation between RCTs and policy making, suggest that advocates of RCTs can help institutions more by putting 'the evidence from trials in its proper context, clarify[ing] the conditions under which interventions work or do not work and why' (p. 398). Such concerns about RCTs resonate with Pawson's (2013) comment that an intervention would be well advised to 'better implement it through A, B, C … better to target it at D, E, F … and better beware of the pitfalls of G, H, I' (p. 190). In other words, in addition to needing to know 'how much' effect a treatment has, and on whom, educational researchers also need to know something that RCTs generally do not indicate, which is *why* some students do or do not respond to an intervention, why others have an excessive response and why others experience side effects or adverse effects, i.e. why there is such variability in effects (Morrison, 2001).

The ethical dimension

Whilst we address ethical aspects of experiments later in this chapter, at this point we suggest that, in order to judge whether an experiment has or has not 'worked', and for whom, it is important to consider the possible fallout from them. Here, for an experiment to 'work' in ethical terms, it should ensure that it has brought no negative or harmful (e.g. psychological, physical, social, emotional) direct or indirect side effects. For example, an experiment to improve student performance in mathematics may succeed in raising performance levels, but at the cost of demotivating students, putting them under immense pressure and turning them off mathematics for life. This is a problem encountered in the 'shadow side' of school (Bray and Lykins, 2012), where students attend private tutorial centres and work with private tutors to improve their test scores in highly competitive systems of schooling; their performance might increase, but so might their dislike of the subject and their anxiety and stress levels.

In approaching any conclusion that the experiment has 'worked', researchers will need to demonstrate that:

- design protocols have been followed;
- randomization has been used appropriately;
- the sample size is suitable;
- the statistical power of the test is appropriate;
- suitable controls have been in place in the experiment;
- extraneous factors have been excluded;
- threats to internal and external validity have been addressed;
- reliability has been addressed;
- appropriate pre-tests and post-tests have been applied, for example, not too easy, not too difficult and with suitable item discriminability (see Chapter 27);
- appropriate proxy measures have been used;
- the correct units of analysis have been used (e.g. individual or cluster analysis);
- appropriate metrics have been used;
- appropriate statistics have been used;
- appropriate criteria for judging 'effectiveness' have been used;
- ethical issues have been addressed.

In noting the affinity between RCTs in education and clinical trials (cf. Torgerson, 2009), responding to the need for recognizing the importance of detail, contingencies and contexts, we suggest that educational research could benefit from the rigour attached to RCTs in medicine, as pharmacopeias indicate: whether a medicine is freely available or a controlled drug; dosage strengths, frequency, quantities and outcomes (dose-response testing); patient screening and diagnosis; security, safety and misuse; indications; contraindications; side effects and adverse effects; delayed effects; register of providers and users; treatment regimens; cautions; patients at risk (e.g. by age, abnormality, special features); presence of other illnesses and other medicines (comorbidities); and methods of treatment. The equivalence of these in RCTs in education is currently difficult to see in their planning, design, conduct, analysis and reporting.

Interventions in experiments in education must take account of a host of factors, contexts and systems in which they exist; rather than trying to control out such factors, contexts and systems, they occupy a central position. In educational research, RCTs have an important place, but theirs is not the entire story of 'what works' when considering the whole system of people, contingencies, changes, contexts, education systems and policy making (cf. Pearce and Raman, 2014) which

obtain in a dynamic, non-linear, interconnected system such as education (Morrison, 2012). Supplementary and complementary methods and data may be useful here.

20.3 Designs in educational experiments

There are several different kinds of experimental design, such as (e.g. Denscombe, 2014):

- the controlled experiment in laboratory conditions (the 'true' experiment): two or more groups;
- the randomized controlled trial;
- the field or quasi-experiment (in the natural setting rather than the laboratory, but where variables are isolated, controlled and manipulated);
- the natural experiment (in which it is not possible to isolate and control variables);
- the retrospective experiment (where the researcher moves from an observed effect and tests to find the likely cause (*ex post facto* research)).

The laboratory experiment (the classic 'true' experiment) is conducted in a specially contrived, artificial environment, so that variables can be isolated, controlled and manipulated (as in the example of the wheat seeds earlier). However, schools and classrooms are not the antiseptic, reductionist, analysed-out or analysable-out world of the laboratory. Indeed the successionist conceptualization of causality (Harré, 1972), wherein researchers make inferences about causality on the basis of observation, must admit its limitations. It is dangerous to infer causes from effects or multiple causes from multiple effects. Generalizability from the laboratory to the classroom is dangerous, yet with field experiments, with their loss of control of variables, generalizability might be equally dangerous.

Sometimes it is not possible, desirable or ethical to set up a laboratory or field experiment. For example, let us imagine that we wanted to investigate the trauma effects on people in road traffic accidents. We could not require a participant to run under a bus, or another to stand in the way of a moving lorry, or another to be hit by a bicycle, and so on. Instead we might examine hospital records to see the trauma effects of victims of bus accidents, lorry accidents and bicycle accidents, and see which group seems to have sustained the greatest traumas. It may be that the lorry accident victims had the greatest trauma, followed by the bus victims, followed by the bicycle victims. Now, although it is not possible to say with 100 per cent certainty what caused the trauma, one could make an intelligent guess that those involved in lorry accidents suffer the worst injuries. Here we look at the outcomes and work backwards to examine possible causes, i.e. we can come to some likely defensible conclusions.

Frequently in experiments on learning in classroom settings the independent variable is a stimulus of some kind, for example, a new method in arithmetical computation, and the dependent variable is a response, for example, the time taken to do twenty sums using the new method. Most empirical studies in educational settings, however, are quasi-experimental rather than experimental. Important differences between the quasi-experiment and the true experiment are that the randomization and controls operating in the true experiment are only partially present, or indeed completely absent, in the quasi-experiment, for example, the groups in the experiment may have been constituted by means other than random selection, or some of the isolation and control of variables may be impossible. In this chapter we identify the essential features of true experimental and quasi-experimental designs, our intention being to introduce the reader to the meaning and purpose of control in educational experimentation.

In experiments, researchers can remain relatively aloof from the participants, bringing a degree of objectivity to the research (Robson, 2002, p. 98). Observer effects can distort the experiment: for example, researchers may record inconsistently, or inaccurately or selectively, or, less consciously, they may be having an effect on the experiment (the problem of bias, deliberate or unconscious). Further, participant effects might distort the experiment (see the discussion of the Hawthorne effect in Chapter 14); the fact of simply being in an experiment, rather than what the experiment is doing, might be sufficient to alter participants' behaviour.

In medical experiments these twin concerns are addressed by having experiments which are blind or double blind and by giving placebos to certain participants, to monitor any changes. In blind experiments, participants are not told whether they are in a control group or an experimental group, though which they are is known to the researcher. In a double-blind experiment not even the researcher knows whether a participant is in the control or experimental group; that knowledge resides with a third party. These are intended to reduce the subtle effects of participants knowing whether they are in a control or experimental group. In educational research it is easier to conduct a blind experiment than a double-blind experiment, and it is even possible not to tell participants that they are in an experiment at all, or to tell them that the experiment is about X when, in fact, it is about Y, i.e. to 'put them off the scent'. This form of deception needs to be justified; a common justification is that it enables the

experiment to be conducted under more natural conditions, without participants altering their everyday behaviour.

In the outline of research designs that follows, we use symbols and conventions from Campbell and Stanley (1963):

- *X* represents the exposure of a group to an experimental variable or event, the effects of which are to be measured;
- *O* refers to the process of observation or measurement;
- *X*s and *O*s in a given row are applied to the same persons;
- left to right order indicates temporal sequence;
- *X*s and *O*s vertical to one another are simultaneous;
- *R* indicates random assignment to separate treatment groups;
- parallel rows unseparated by dashes represent comparison groups equated by randomization, while those separated by a dashed line represent groups not equated by random assignment.

20.4 True experimental designs

There are several variants of the 'true' experimental design, and we consider many of these below:

- the pre-test–post-test control and experimental group design;
- the two control groups and one experimental group pre-test–post-test design;
- the post-test control and experimental group design;
- the post-test two experimental groups design;
- the pre-test–post-test two treatment design;
- the matched pairs design;
- the factorial design;
- the parametric design;
- repeated measures designs.

The laboratory experiment typically has to identify and control a large number of variables, and this may not be possible in education. Further, the laboratory environment itself can have an effect on the experiment, or it may take some time for a particular intervention to manifest its effects (e.g. a particular reading intervention may have little immediate effect but may have a delayed effect in promoting a liking for reading in adult life, or may have a cumulative effect over time).

A 'true' experiment includes several key features:

- one or more control groups;
- one or more experimental groups;
- random sampling from a population;
- random allocation to control and experimental groups;
- pre-test of the groups to ensure parity;
- one or more interventions to the experimental group(s);
- isolation, control and manipulation of independent variables;
- post-test of the groups to see the effects on the dependent variable;
- post-test of the groups to see the effects on the groups;
- non-contamination between the control and experimental groups.

If an experiment does not possess all of these features then it is a quasi-experiment: it may look *as if* it is an experiment ('quasi' means 'as if') but it is not a true experiment, only a variant on it.

An alternative to the laboratory experiment is the quasi-experiment or field experiment, including:

- the one-group pre-test–post-test;
- the non-equivalent control group design;
- the time series design.

We consider these below. Field experiments have less control over experimental conditions or extraneous variables than a laboratory experiment, and, hence, inferring causality is more contestable, but they have the attraction of taking place in a natural setting. Extraneous variables may include:

- participant factors (they may differ on important characteristics between the control and experimental groups);
- intervention factors (the intervention may not be exactly the same for all participants, varying, for example, in sequence, duration, degree of intervention and assistance, and other practices and contents);
- situational factors (the experimental conditions may differ).

These can lead to experimental error, in which the results may not be due to the independent variables in question. Ary *et al*. (2006) and Shadish *et al*. (2002) provide a useful overview of true and quasi experiments.

The pre-test–post-test control and experimental group design

A complete exposition of experimental designs is beyond the scope of this chapter. In the brief outline

that follows, we have selected one design from the comprehensive treatment of the subject by Campbell and Stanley (1963) in order to identify the essential features of what they term a 'true experimental' and what Kerlinger (1970) refers to as a 'good' design. Along with its variants, the chosen design is commonly used in educational experimentation (e.g. Schellenberg, 2004).

The pre-test–post-test control group design can be represented as:

Experimental	RO_1	X	O_2
Control	RO_3		O_4

The two control groups and one experimental group pre-test–post-test design

This is the Solomon design, intended to identify the interaction effect that may occur if the subject deduces the desired result from looking at the pre-test and the post-test. It is the same as the RCT above, except that there are two control groups instead of one. In the standard RCT, any change in the experimental group can be due to the intervention or the pre-test, and any change in the control group can be due to the pre-test. In the Solomon variant the second control group receives the intervention but no pre-test. This can be modelled thus:

Experimental	RO_1	X	O_2
$Control_1$	RO_3		O_4
$Control_2$		X	O_5

Thus any change in this second control group can only be due to the intervention. A variant of the Solomon three-group design is the Solomon four-group design (with one experimental group and three control groups). We refer readers to Bailey (1994, pp. 231–4), Ary *et al*. (2009) and Shadish *et al*. (2002) for a full explication of this technique and its variants.

The post-test control and experimental group design

Here participants are randomly assigned to a control group and an experimental group, but there is no pre-test. The experimental group receives the intervention and the two groups are given only a post-test. The design is:

Experimental	R_1	X	O_1
Control	R_2		O_2

The post-test two experimental groups design

Here participants are randomly assigned to each of two experimental groups. Experimental group 1 receives intervention 1 and experimental group 2 receives intervention 2. Only post-tests are conducted on the two groups. The design is:

$Experimental_1$	R_1	X_1	O_1
$Experimental_2$	R_2	X_2	O_2

The pre-test–post-test two treatment design

Here participants are randomly allocated to each of two experimental groups. Experimental group 1 receives intervention 1 and experimental group 2 receives intervention 2. Pre-tests and post-tests are conducted to measure changes in individuals in the two groups. The design is:

$Experimental_1$	RO_1	X_1	O_2
$Experimental_2$	RO_3	X_2	O_4

The true experiment can also be conducted with one control group and two or more experimental groups. So, for example, the design might be:

$Experimental_1$	RO_1	X_1	O_2
$Experimental_2$	RO_3	X_2	O_4
Control	RO_5		O_6

This can be extended to the post-test control and experimental group design and the post-test two experimental groups design, and the pre-test–post-test two treatment design.

The matched pairs design

As the name suggests, here participants are allocated to control and experimental groups randomly, but the basis of the allocation is that one member of the control group is matched to a member of the experimental group on the several independent variables considered important for the study (e.g. those independent variables that are considered to have an influence on the dependent variable, such as sex, age, ability). So, first, pairs of participants are selected who are matched in terms of the independent variable under consideration (e.g. whose scores on a particular measure are the same or similar), and then each one of the pair is randomly assigned to the control or experimental group. Randomization takes place at the pair rather than the group level. Though, as its name suggests, this ensures

effective matching of control and experimental groups, in practice it may not be easy to find sufficiently close matching, particularly in a field experiment, though finding such a close match in a field experiment may increase the control of the experiment considerably. Matched pairs designs are useful if the researcher cannot be certain that individual differences will not obscure treatment effects, as it enables these individual differences to be controlled.

Borg and Gall (1979, p. 547) set out a useful series of steps in the planning and conduct of a matched pairs experiment:

Step 1: Carry out a measure of the dependent variable.
Step 2: Assign participants to matched pairs, based on the scores and measures established from Step 1.
Step 3: Randomly assign one person from each pair to the control group and the other to the experimental group.
Step 4: Administer the experimental treatment/intervention to the experimental group and, if appropriate, a placebo to the control group. Ensure that the control group is not subject to the intervention.
Step 5: Carry out a measure of the dependent variable with both groups and compare/measure them in order to determine the effect and its size on the dependent variable.

Borg and Gall indicate that difficulties arise in the close matching of the sample of the control and experimental groups. This involves careful identification of the variables on which the matching must take place. They suggest (p. 547) that matching on a number of variables that correlate with the dependent variable is more likely to reduce errors than matching on a single variable. The problem is that the greater the number of variables that have to be matched, the harder it is actually to find the sample of people who are matched. Hence the balance must be struck between having too few variables such that error can occur, and having so many variables that it is impossible to draw a sample. Instead of matched pairs, random allocation is possible, and this is discussed below.

Mitchell and Jolley (1988, p. 103) pose three important questions that researchers need to consider when comparing two groups:

- Are the two groups equal at the commencement of the experiment?
- Would the two groups have grown apart naturally, regardless of the intervention?
- To what extent has initial measurement error of the two groups been a contributory factor in differences between scores?

Borg and Gall (1979) draw attention to the need to specify the degree of exactitude (or variance) of the match. For example, if the subjects were to be matched on, say, linguistic ability as measured in a standardized test, it is important to define the limits of variability that will be used to define the matching (e.g. ±3 points). As before, the greater the degree of precision in the matching here, the closer will be the match, but the greater the degree of precision the harder it will be to find an exactly matched sample.

One way of addressing precision is to place all the subjects in rank order on the basis of the scores or measures of the dependent variable. Then the first two subjects become one matched pair (in which one is allocated to the control group and one to the experimental group randomly, e.g. by tossing a coin), subjects three and four become the next matched pair, subjects five and six become the next matched pair, and so on until the sample is drawn. Here the loss of precision is counterbalanced by the avoidance of the loss of subjects.

The alternative to matching that has been discussed earlier in the chapter is randomization. Smith (1991, p. 215) suggests that matching is most widely used in quasi-experimental and non-experimental research, and is a far inferior means of ruling out alternative causal explanations than randomization.

The factorial design

In an experiment there may be two or more independent variables acting on the dependent variable. For example, performance in an examination may be a consequence of availability of resources (independent variable one: limited availability, moderate availability, high availability) and motivation for the subject studied (independent variable two: little motivation, moderate motivation, high motivation). Each independent variable is studied at each of its levels (in the example here it is three levels for each independent variable). Participants are randomly assigned to groups that cover all the possible combinations of levels of each independent variable, for example:

Independent variable	*Level 1*	*Level 2*	*Level 3*
Availability of resources	limited availability (1)	moderate availability (2)	high availability (3)
Motivation for the subject studied	little motivation (4)	moderate motivation (5)	high motivation (6)

Here the possible combinations are: 1+4, 1+5, 1+6, 2+4, 2+5, 2+6, 3+4, 3+5 and 3+6. This yields

nine groups (3 × 3 combinations). Pre-tests and post-tests or post-tests only can be conducted. It might show, for example, that limited availability of resources and little motivation had a large influence on examination performance, whereas moderate and high availability of resources did not, or that high availability and high motivation had a large effect on performance, whereas high motivation and limited availability did not, and so on.

This example assumes that there are the same numbers of levels for each independent variable; however, this may not be the case. One variable may have, say, two levels, another three levels and another four levels. Here the possible combinations are $2 \times 3 \times 4 = 24$ levels and, therefore, 24 experimental groups. One can see that factorial designs quickly generate several groups of participants. A common example is a 2 × 2 design, in which two independent variables each have two values (i.e. four groups). Here experimental group 1 receives the intervention with independent variable 1 at level 1 and independent variable 2 at level 1; experimental group 2 receives the intervention with independent variable 1 at level 1 and independent variable 2 at level 2; experimental group 3 receives the intervention with independent variable 1 at level 2 and independent variable 2 at level 1; experimental group 4 receives the intervention with independent variable 1 at level 2 and independent variable 2 at level 2.

Factorial designs also have to take account of the interaction of the independent variables. For example one factor (independent variable) may be 'sex' and the other 'age' (Figure 20.2). The researcher may be investigating their effects on motivation for learning mathematics.

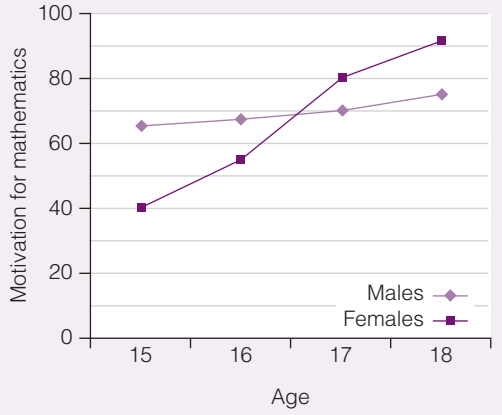


FIGURE 20.2 Interaction effects in an experiment

In Figure 20.2 the difference in motivation for mathematics is not constant between males and females; it varies according to the age of the participants. There is an interaction effect between age and sex, such that the effect of sex depends on age. A factorial design is useful for examining interaction effects.

At their simplest, factorial designs may have two levels of an independent variable, for example, its presence or absence, but, as has been seen here, it can quickly become more complex. That complexity is bought at the price of increasing exponentially the number of groups required.

The parametric design

Here participants are randomly assigned to groups whose parameters are fixed in terms of the levels of the independent variable that each receives. For example, let us imagine that an experiment is conducted to improve the reading abilities of poor, average, good and outstanding readers (four levels of the independent variable 'reading ability'). Four experimental groups are set up to receive the intervention, thus: experimental group 1 (poor readers); experimental group 2 (average readers), experimental group 3 (good readers) and experimental group 4 (outstanding readers). The control group (group 5) would receive no intervention. The researcher could chart the differential effects of the intervention on the groups, and thus have a more sensitive indication of its effects than if there was only one experimental group containing a wide range of reading abilities; the researcher would know which group was most and least affected by the intervention. Parametric designs are useful if an independent variable is considered to have different levels or a range of values which may have a bearing on the outcome (confirmatory research) or if the researcher wishes to discover whether different levels of an independent variable have an effect on the outcome (exploratory research).

Repeated measures designs

Here participants in the experimental groups are tested under two or more experimental conditions. So, for example, a member of the experimental group may receive more than one 'intervention', which may or may not include a control condition. This offers considerable potential for control, as it is exactly the same person receiving different interventions. Order effects raise their heads here: the order in which the interventions are sequenced may have an effect on the outcome; the first intervention may have an influence – a carry-over effect – on the second, and the second intervention may have an influence on the third, and so on. Further, early interventions may have a greater effect than later

interventions. To overcome this it is possible to randomize the order of the interventions and assign participants randomly to different sequences, though this may not ensure a balanced sequence. Rather, a deliberate ordering may have to be planned, for example, in a three-intervention experiment:

Group 1 receives intervention 1 followed by intervention 2, followed by intervention 3;
Group 2 receives intervention 2 followed by intervention 3, followed by intervention 1;
Group 3 receives intervention 3 followed by intervention 1, followed by intervention 2;
Group 4 receives intervention 1 followed by intervention 3, followed by intervention 2;
Group 5 receives intervention 2 followed by intervention 1, followed by intervention 3;
Group 6 receives intervention 3 followed by intervention 2, followed by intervention 1.

Repeated measures designs are useful if it is considered that order effects are either unimportant or unlikely (see Figure 20.3), or if the researcher cannot be certain that individual differences will not obscure treatment effects, as it enables these individual differences to be controlled.

20.5 Quasi-experimental designs

Often in educational research, it is simply not possible for investigators to undertake true experiments, for example, random selection and random assignment of participants to control or experimental groups. Quasi-experiments are the stuff of field experimentation, i.e. outside the laboratory. At best, they may be able to employ something approaching a true experimental design in which they have control over what Campbell and Stanley (1963) refer to as 'the who and to whom of measurement', but lack control over 'the when and to whom of exposure' or the randomization of exposures – essential if true experimentation is to take place. These situations are quasi-experimental and the methodologies employed by researchers are termed quasi-experimental designs. (Kerlinger (1970) refers to quasi-experimental situations as 'compromise designs', an apt description when applied to much educational research where the random selection or random assignment of schools and classrooms is quite impracticable.)

Quasi-experiments come in several forms, for example:

- pre-experimental designs: the one-group pre-test–post-test design; the one-group post-tests only design; the non-equivalent post-test only design;
- pre-test–post-test non-equivalent group design;
- one-group time series.

We consider these below.

A pre-experimental design: the one-group pre-test–post-test

A pre-experimental design is so named because it offers little or even no control over extraneous variables (Ary *et al*., 2009). Very often, reports about the value of a new teaching method or interest aroused by a curriculum innovation reveal that a researcher has measured a group on a dependent variable (O_1), for example, attitudes towards minority groups, and then introduced an

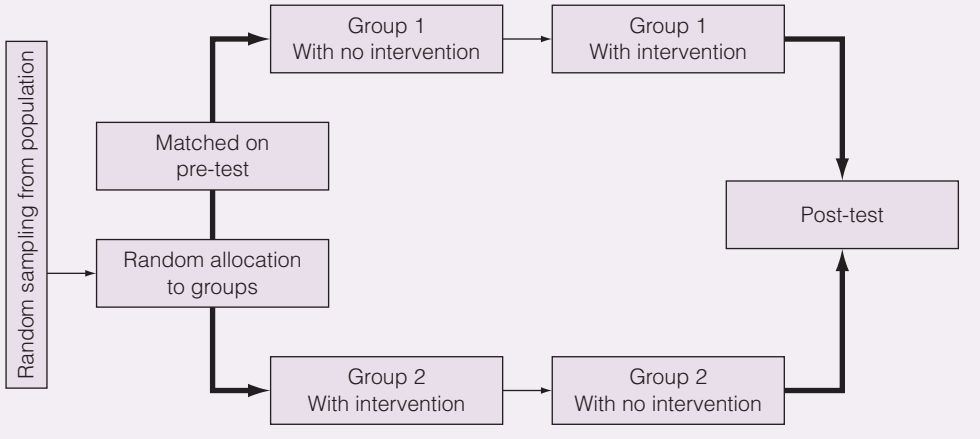


FIGURE 20.3 Two groups receiving both conditions (repeated measures)

experimental manipulation (X), perhaps a ten-week curriculum project designed to increase tolerance of ethnic minorities. Following the experimental treatment, the researcher has again measured group attitudes (O_2) and proceeded to account for differences between pre-test and post-test scores by reference to the effects of X.

The one-group pre-test–post-test design can be represented as:

Experimental	O_1	X	O_2

Suppose that just such a project has been undertaken and that the researcher finds that O_2 scores indicate greater tolerance of ethnic minorities than O_1 scores. How justified is she in attributing the cause of such differences to the experimental treatment (X), that is, the term's project work? At first glance the assumption of causality seems reasonable enough. The situation is not that simple, however. Compare for a moment the circumstances represented in our hypothetical educational example with those which typically obtain in experiments in the physical sciences. A physicist who applies heat to a metal bar can confidently attribute the observed expansion to the rise in temperature that she has introduced because within the confines of her laboratory she has excluded (i.e. controlled) all other extraneous sources of variation. The same degree of control can never be attained in educational experimentation. At this point readers may care to reflect upon some possible influences other than the ten-week curriculum project that might account for the differences in our hypothetical educational example.

They may conclude that factors to do with the pupils, the teacher, the school, the classroom organization, the curriculum materials and their presentation, how the subjects' attitudes were measured, to say nothing of the thousand and one other events that occurred in and about the school during the course of the term's work, might all have exerted some influence upon the observed differences in attitude. These kinds of extraneous variables which are outside the experimenter's control in one-group pre-test–post-test designs threaten to invalidate their research efforts. We later identify a number of such threats to the validity of educational experimentation.

A pre-experimental design: the one-group post-tests only design

Here an experimental group receives the intervention and then takes the post-test. Though this has some features of an experiment (an intervention and a post-test), the lack of a pre-test, of a control group, of random allocation and of controls renders this a flawed methodology.

A pre-experimental design: the post-tests only non-equivalent groups design

Again, though this appears to be akin to an experiment, the lack of a pre-test, of matched groups, of random allocation and of controls renders this a flawed methodology.

A quasi-experimental design: the pre-test–post-test non-equivalent group design

One of the most commonly used quasi-experimental designs in educational research can be represented as:

Experimental	O_1	X	O_2
Control	O_3		O_4

The dashed line separating the parallel rows in the diagram of the non-equivalent control group indicates that the experimental and control groups have not been equated by randomization – hence the term 'non-equivalent'. The addition of a control group makes the present design a decided improvement over the one-group pre-test–post-test design, as, to the degree that experimenters can make experimental and control groups as equivalent as possible, they can avoid the equivocality of interpretations that plague the pre-experimental design discussed earlier. The equivalence of groups can be strengthened by matching, followed by random assignment to experimental and control treatments.

Where matching is not possible, the researcher is advised to use samples from the same population or samples that are as alike as possible (Kerlinger, 1970). Where intact groups differ substantially, however, matching is unsatisfactory due to regression effects which lead to different group means on post-test measures.

The one-group time series

Here the one group is the experimental group, and it is given more than one pre-test and more than one post-test. The time series uses repeated tests or observations both before and after the treatment, which, in effect, enables the participants to become their own controls, which reduces the effects of reactivity. Time series allow for trends to be observed, and avoids reliance on only one single pre-testing and post-testing data-collection point. This enables trends to be observed such as: no effect at all (e.g. continuing an existing upward, downward or even trend), a clear effect (e.g. a sustained rise or drop in performance), delayed effects (e.g. some time after the intervention has occurred). Time series studies have the potential to increase reliability.

20.6 Single-case ABAB design

At the beginning of Chapter 19, we described case study researchers as typically engaged in observing the characteristics of an individual unit, be it a child, a classroom, a school, or a whole community. We went on to contrast case study researchers with experimenters whom we described as typically concerned with the manipulation of variables in order to determine their causal significance. That distinction, as we shall see, is only partly true.

Increasingly, in recent years, single-case research as an experimental methodology has extended to such diverse fields as clinical psychology, medicine, education, social work, psychiatry and counselling. Most of the single-case studies carried out in these (and other) areas share the following characteristics:

- they involve the continuous assessment of some aspect of human behaviour over a period of time, requiring on the part of the researcher the administration of measures on multiple occasions within separate phases of a study;
- they involve 'intervention effects' which are replicated in the same subject(s) over time.

Continuous assessment measures are used as a basis for drawing inferences about the effectiveness of intervention procedures.

The characteristics of single-case research studies are discussed by Kazdin (1982) and Ary *et al*. (2002) in terms of ABAB designs, the basic experimental format in most single-case research. ABAB designs consist of a family of procedures in which observations of performance are made over time for a given client or group of clients. Over the course of the investigation, changes are made in the experimental conditions to which the client is exposed. The basic rationale of the ABAB design is illustrated in Figure 20.4. What it does is this. It examines the effects of an intervention by alternating the baseline condition (the A phase), when no intervention is in effect, with the intervention condition (the B phase). The A and B phases are then repeated to complete the four phases. As Kazdin and Ary *et al*. note, the effects of the intervention are clear if performance improves during the first intervention phase, reverts to or approaches original baseline levels of performance when the treatment is withdrawn, and improves again when treatment is recommenced in the second intervention phase.

An example of the application of the ABAB design in an educational setting is provided by Dietz (1977), whose single-case study sought to measure the effect that a teacher could have upon the disruptive behaviour of an adolescent boy whose persistent talking disturbed his fellow classmates in a special education class.

In order to decrease the unwelcome behaviour, a reinforcement programme was devised in which the boy could earn extra time with the teacher by decreasing the number of times he called out. The boy was told that when he made three (or fewer) interruptions during

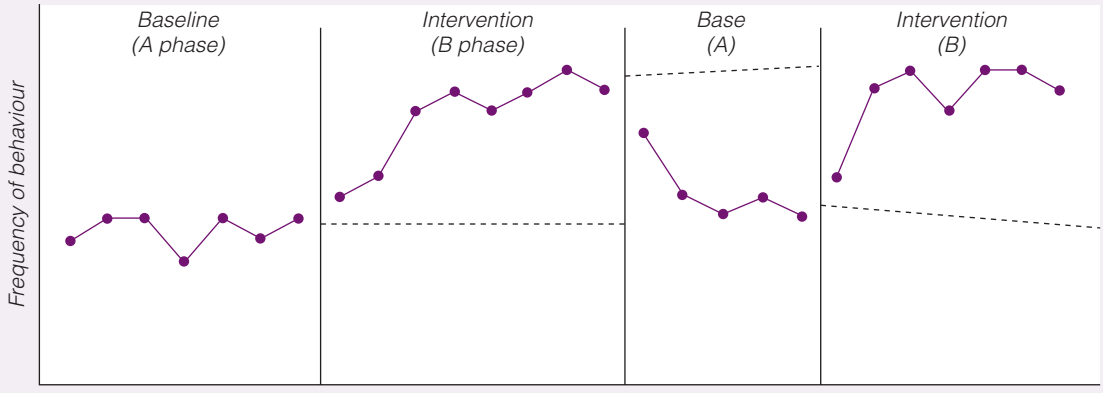


FIGURE 20.4 The ABAB design
Source: Adapted from Kazdin (1982)

any fifty-five-minute class period, the teacher would spend extra time working with him. In the technical language of behaviour modification theory, the pupil would receive reinforcing consequences when he was able to show a low rate of disruptive behaviour (in Figure 20.5 this is referred to as 'differential reinforcement of low rates' or DRL).

When the boy was able to desist from talking aloud on fewer than three occasions during any timetabled period, he was rewarded by the teacher spending fifteen minutes with him helping him with his learning tasks. The pattern of results displayed in Figure 20.5 shows the considerable changes that occurred in the boy's behaviour when the intervention procedures were carried out and the substantial increases in disruptions towards baseline levels when the teacher's rewarding strategies were withdrawn. Finally, when the intervention was reinstated, the boy's behaviour is seen to improve again.

Ary *et al*. (2002) provide an example of an ABAB design with a single case of an eight-year-old boy who was developmentally disabled. There is also the famous example of the 'still face experiment' with young babies(e.g.www.youtube.com/watch?v=apzXGEbZht0) in which a mother interacts positively with the baby for some time, then adopts an expressionless, unresponsive 'still face', and repeats this sequence, and we are able to observe the baby's increasingly frantic attempts to attract the mother's attention.

The single-case research design is uniquely able to provide an experimental technique for evaluating interventions for the individual subject. Moreover, such interventions can be directed towards the particular subject or group and replicated over time or across behaviours, situations or persons. Single-case research offers an alternative strategy to the more usual methodologies based on between-group designs. There are, however, a number of problems that arise in connection with the use of single-case designs having to do with ambiguities introduced by trends and variations in baseline phase data and with the generalizability of results from single-case research.

20.7 Procedures in conducting experimental research

An experimental investigation must follow a set of logical procedures. Those that we now enumerate, however, should be treated with some circumspection. It is extraordinarily difficult (and foolhardy) to lay down clear-cut rules as guides to experimental research. At best, we can identify an ideal route to be followed, mindful that educational research rarely proceeds in such a systematic fashion.

First, the researcher must identify and define the research problem as precisely as possible, always supposing that the problem is amenable to experimental methods.

Second, she must formulate hypotheses that she wishes to test. This involves making predictions about relationships between specific variables and at the same time making decisions about other variables that are to

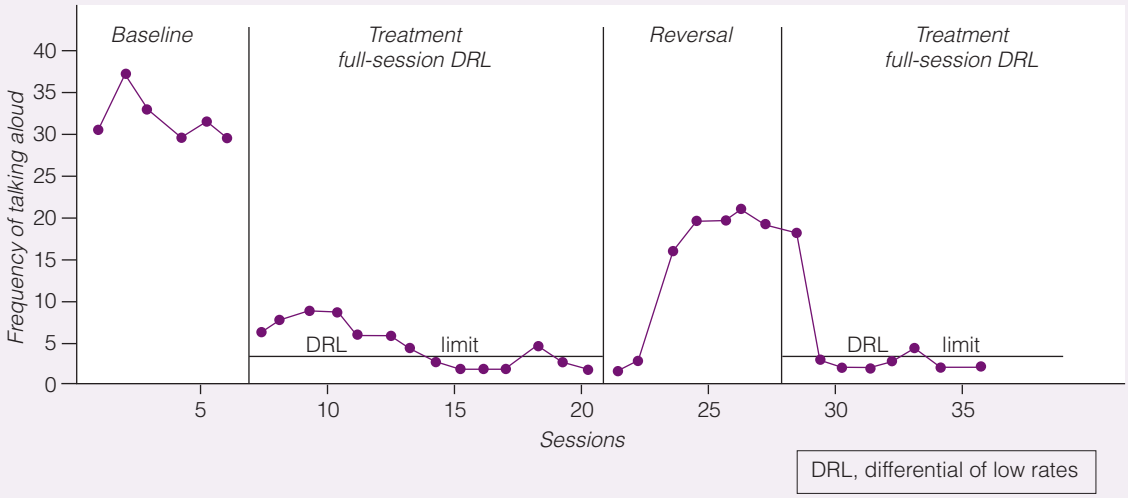


FIGURE 20.5 An ABAB design in an educational setting
Source: Kazdin (1982)

be excluded from the experiment by means of controls. Variables, remember, must have two properties. First, they must be measurable. Physical fitness, for example, is not directly measurable until it has been operationally defined. Making the variable 'physical fitness' operational means simply defining it by letting something else that is measurable stand for it – a gymnastics test, perhaps (a proxy variable). Second, the proxy variable must be a valid indicator of the hypothetical variable in which one is interested. That is to say, a gymnastics test probably is a reasonable proxy for physical fitness; height, on the other hand, most certainly is not. Excluding variables from the experiment is inevitable, given constraints of time and money. It follows therefore that one must set up priorities among the variables in which one is interested so that the most important of them can be varied experimentally whilst others are held constant.

Third, the researcher must select appropriate levels at which to test the independent variables. By way of example, suppose an educational psychologist wishes to find out whether longer or shorter periods of reading make for reading attainment in school settings (see Simon, 1978). She will hardly select five-hour and five-minute periods as appropriate levels; rather, she is more likely to choose thirty-minute and sixty-minute levels, in order to compare with the usual timetabled periods of forty-five minutes' duration. In other words, the experimenter will vary the stimuli at such levels as are of practical interest in the real-life situation. Pursuing the example of reading attainment further, our hypothetical experimenter will be wise to vary the stimuli in large enough intervals so as to obtain measurable results. Comparing reading periods of forty-four minutes or forty-six minutes with timetabled reading lessons of forty-five minutes is scarcely likely to result in observable differences in attainment.

Similarly Torgerson and Torgerson (2008) alert researchers to 'ceiling and floor effects' (pp. 147–8). A 'ceiling effect' is where a test is too easy for the participants, whilst a 'floor effect' is where it is too difficult. This rehearses the need not only to pilot the test but to ensure that item discriminability and appropriate scaling have been addressed (see Chapter 27 of the present volume). The authors note that if there is a ceiling or floor effect then it may lead to the false conclusion that an intervention has not worked.

Fourth, the researcher must decide which kind of experiment she will adopt, perhaps from the varieties set out in this chapter.

Fifth, in planning the design of the experiment, the researcher must take account of the population to which she wishes to generalize her results. This involves her in decisions over sample sizes, sampling methods and contextual matters. Sampling decisions may include questions of funds, staffing and the amount of time available for experimentation. However, one general rule of thumb is to try to make the sample as large as possible so that even small effects can reveal themselves which might otherwise be lost with small samples, even though the trade-off here is that, with large samples, it is easier to achieve statistical significance (i.e. it is easier to find a statistically significant difference between the control group and the experimental group) than it is with a small sample (statistical significance being, in part, a function of sample size) (cf. Torgerson and Torgerson, 2008, p. 128), though measures of effect size overcome this problem. Further, it is important, where possible, to use a random, probability sample, as this not only permits a greater range of statistics to be used (e.g. t-tests and Analysis of Variance (ANOVA), both of which are important in experiments, see Chapter 41), but it also enables the findings to have greater generalizability (external validity), i.e. to represent the wider population. Contextual similarity also has to be considered in addressing generalizability, as results in one context or culture, regardless of statistical significance and effect size, may not travel well to a very different context or culture (Cartwright and Hardie, 2012).

Sixth, with problems of validity in mind, the researcher must select instruments, choose tests and decide upon appropriate methods of analysis (typically t-tests and measures of effect size are used to determine whether there are any statistically significant or sizeable differences that are worthy of note, respectively, between the control and experimental groups).

Seventh, before embarking upon the actual experiment, the researcher must pilot the experimental procedures and measures to identify possible problems in connection with any aspect of the investigation. This is of crucial importance.

Eighth, during the experiment itself, the researcher must endeavour to follow tested and agreed-on procedures to the letter (standard protocols). The standardization of instructions and adherence to them, the exact timing of experimental sequences and the meticulous recording and checking of observations are all the hallmark of the competent researcher.

With her data collected, the researcher faces the most important part of the whole enterprise. Processing data, analysing results and drafting reports are all demanding activities, both in intellectual effort and time. Often this last part of the experimental research is given too little time in the overall planning of the investigation. Experienced researchers rarely make such a

mistake; unanticipated disasters teach the hard lesson of leaving ample time for the analysis and interpretation of experimental findings.

We suggest a ten-step model for the conduct of the experiment:

Step 1: Identify the purpose of the experiment.
Step 2: Select the relevant variables.
Step 3: Specify the level(s) of the intervention (e.g. low, medium, high intervention).
Step 4: Isolate and control the experimental conditions and environment.
Step 5: Select the appropriate experimental design.
Step 6: Administer the pre-test.
Step 7: Sample the relevant population and assign the participants to the groups.
Step 8: Conduct the intervention.
Step 9: Conduct the post-test.
Step 10: Analyse the results.

The sequence of steps 6 and 7 can be reversed; the intention in putting them in the present sequence is to ensure that the two groups are randomly selected, allocated and matched. In calculating differences or similarity between groups at the stages of the pre-test and the post-test, the t-test for independent samples or ANOVA are often used.

20.8 Threats to internal and external validity in experiments

Chapter 14 indicated several threats to the internal and external validity of experiments, and we refer the reader to this chapter. In that chapter threats to internal validity (the validity of the research design, process, instrumentation and measurement) were seen to reside in:

- history
- maturation
- statistical regression
- testing
- instrumentation
- selection
- experimental mortality
- instrument reactivity
- selection–maturation interaction
- Type I and Type II errors.

To this, Hammersley (2008, p. 4) adds the point that not all the confounding variables may be properly controlled in the randomization process.

In Chapter 14, too, threats to external validity (wider generalizability) were seen to reside in:

- failure to describe independent variables explicitly;
- lack of representativeness of available and target populations;
- the Hawthorne effect;
- inadequate operationalizing of dependent variables;
- sensitization/reactivity to experimental/research conditions;
- interaction effects of extraneous factors and experimental/research treatments;
- invalidity or unreliability of instruments;
- ecological validity;
- multiple treatment validity.

To this, Hammersley (2008, p. 4) adds the point that, in principle, a laboratory trial, in which variables are controlled, misrepresents the 'real' world of the classroom or school in which the variables are far less controlled, i.e. the findings may not be transferable to wider conditions and situations. Further, Torgerson (2009) notes that, unlike crops in agriculture (the origin of Fisher's (1966) experimental model), humans do not always act as the experimenter would like or in ways in which the experimenter has predicted (p. 315) (see also Camburn *et al*., 2015, p. 8).

One can add to these factors the matter that statistical significance can be found comparatively easily if sample sizes are large (Kline, 2004) (hence the need to consider placing greater reliance on effect size rather than statistical significance, discussed in Chapter 39). Further, Torgerson and Torgerson (2003a, 2008) draw attention to the limits of small samples in experimental research, as small samples can fail to spot small effects, thereby risking a Type II error (failing to find an effect when, in fact, it exists: a false negative). As they remark, in a time of evidence-based education and discussions of 'what works', small effects can be useful (Torgerson and Torgerson, 2003a, p. 70), and they give the example where, if a small change in 'delivering' the curriculum leads to improved examination passes of one only child in each class in public examinations, then this could scale up to between 20,000 and 30,000 students across the UK.

Torgerson and Torgerson (2003b, 2008) and Torgerson (2009) also identify several sources of bias in randomized controlled trials, for example:

- use of a *very selective sample* (they give the example of an exclusive girls-only boarding school) and then seeking to generalize the results to a much wider population, for example, an inner-city mixed sex comprehensive (non-selective) school (Torgerson and Torgerson, 2003b, p. 37);

- a *selection bias* (i.e. a non-random selection), if the researcher allocates the students on preference: a non-blind random allocation (Torgerson, 2009, p. 316);
- a *selection bias*, where the experimental group possesses a variable that is related to the outcome variable but which is not included in the intervention (Torgerson and Torgerson, 2003b, pp. 37–8);
- a *dilution bias*, where the control group, not being exposed to the intervention, deliberately seeks out a 'compensating treatment' (p. 38). For example, there may be an experiment to test the effects of increased attention to mathematics in the classroom on mathematics results in public examinations; the control group, not being exposed to what they see as a useful intervention (given that there has to be informed consent), may take private mathematics lessons in order to compensate, thereby disturbing or diluting the findings of the experiment;
- *chance effects*: Torgerson and Torgerson (2003b) give an example of a group of forty children learning spellings, in which four of them were dyslexic, and in which the likelihood of them being randomly allocated to the control group and experimental group evenly (two in each group) was very small, indeed all four could be in one group (either the experimental group or the control group). The researchers argue that this can be addressed through 'minimisation' (p. 40), deliberately ensuring an even split of such students into both groups (e.g. matched pairs allocation);
- *'subversion bias'*, where researchers deliberately breach the requirements of random allocation (hence the need for double-blind experiments or where the researcher is not involved in the randomized allocation);
- *attrition bias*: where some students drop out of the experimental group. (Torgerson and Torgerson (2003b) give the example of students who attend voluntary Saturday morning 'booster classes' and then drop out of the class.) Here, if the researchers had only focused on the results of those students who remained in the Saturday morning classes, then they would have obtained very different results from those which might have been found if the dropouts had not dropped out (e.g. in terms of measured motivation levels and, hence, achievement). There is a risk of 'attrition bias' here (p. 75);
- *reporting or detection bias*: where different researchers or reporters for the control and experimental groups report with differing degrees of detail or inclusion of relevant observations (Torgerson and Torgerson, 2003b, p. 42);
- *exclusion bias*, where members of the experimental group, for reasons other than attrition, do not actually take part in the experiment;
- *marker bias*, if post-tests are marked by researchers who are not blinded with regard to the allocation of participants (Torgerson, 2009, p. 316).

20.9 The timing of the pre-test and the post-test

Experiments often suffer from the problem of only having two time points for measurement: the pre-test and the post-test. It is essential that the researcher plans the timing of the pre-test and the post-test appropriately. Morrison (2009, p. 168) writes that 'experimental procedures are prone to problems of timing – too soon and the effect may not be noticed; too late and the effect might have gone or been submerged by other matters'. The pre-test should be conducted as close to the start of the intervention as possible, to avoid the influence of confounding effects between the pre-test and the start of the intervention; that is quite straightforward.

More difficult is the issue of the timing of the post-test. On the one hand, the argument is strong that it should be as close as possible to the end of the intervention, as this will reduce the possibility of the influence of confounding effects. On the other hand, if it is as close as possible to the end of the intervention it might lead to a false positive, i.e. finding an effect which is transitory or only immediate, i.e. an effect which is not sustained to any worthwhile degree over time. A standard example of this is where an end-of course examination is administered at the last session of the course, or within a week of its completion, and, unsurprisingly perhaps, given the 'recency effect' (in which most recently studied items are more easily recalled than items studied a long time previously), many students score well. However, let us imagine that the post-test (the examination) had been conducted one month later, in which case the students might well have bleached the subject matter from their minds. Or, more problematic in this instance is the familiar case of students revising hard before the post-test (the examination) is administered and they score well, but this time it is not a consequence of the intervention but a rehearsal, practice or revision effect.

It may well be that the effects of a particular intervention may not reveal themselves immediately, but much later. For example, a student may study Shakespeare at age fifteen and, on an outcome measure, may use it to say that she strongly dislikes English literature, but, years later, she may point back to her study of

Shakespeare as sowing the seed for her eventual love of Shakespeare that only developed after she had left school.

On the one hand, too soon the post-test and that effect is lost, it goes unmeasured (and this is a serious problem for the 'what works' movement, which often concerns itself with short-term payback). On the other hand, too long a time lapse and it becomes impossible to determine whether it was a particular independent variable that caused a particular effect, or whether other factors have intervened since the intervention to produce the effect.

One way in which the researcher can overcome the difficulty of the timing of the post-test is to have more than one post-test (e.g. an 'equivalent form' of the post-test, see Chapter 14), with the post-test administered soon after the intervention has ended, and its equivalent form administered after a longer period of time – to determine more long-lasting effects.

20.10 The design experiment

The design experiment can be considered as a special case of a field experiment; it has its roots in experimental research, both in 'true' and quasi-experiments, and is intended to provide formative feedback on, for example, practical problems in, say, teaching and learning, and to bridge the potential gap between research and practice (Brown, 1992, p. 143; Reinking and Bradley, 2008; Bradley and Reinking, 2011; Engeström, 2011; Seel, 2011, p. 925; Anderson and Shattuck, 2012; Laurillard, 2012), in other words, to enhance the external validity of an experiment. The design experiment strives to avoid the artificial world of the laboratory and the lack of applicability to 'real-world problems' that follows from this artificial condition (Bradley and Reinking, 2011; Reinking and Bradley, 2008; Seel, 2011; Laurillard, 2012), and to have direct practical relevance to the complex world of teaching, learning and classrooms. Given their intended direct relevance to classrooms and the field nature – the diverse, complex, 'real world' of an actual classroom – design experiments may not be able to fulfil the requirements of a true experiment, for example, in randomization or in the application of controls. In these respects, design experiments are similar to action research (cf. Anderson and Shattuck, 2012).

Bradley and Reinking (2011), commenting on design experiments in early childhood education, note that they are intended to identify factors within classrooms which promote or inhibit effective teaching and learning and then seek to accentuate the positives and eliminate the negatives. The authors note seven features of design experiments (pp. 312–13):

- they focus on interventions in authentic, real-world settings;
- the role of theory is important in providing a rationale for the intervention, indeed testing the theory is a key purpose of design experiments;
- they have the improvement of practice as their goal, for example, how to improve teaching and learning in authentic settings;
- they are iterative in their data collection, gathering data as the intervention evolves over time and across sites;
- contextual factors influence – both positively and negatively – what happens at the sites of interventions and, hence, the design experiment;
- data collection employs multiple methods;
- they are rooted in pragmatism.

The authors raise six questions that design experiments address:

1. What is the pedagogical goal to be investigated; why is that goal valued and important and what theory and practice and previous empirical work speaks to accomplishing that goal instructionally?
2. What instructional intervention, consistent with a guiding theory, has the potential to achieve the pedagogical goal and why?
3. What factors enhance or inhibit the effectiveness, efficiency and appeal of the instructional intervention in regard to achieving the educational goal?
4. How can the instructional intervention be adapted to achieve the pedagogical goal more effectively and efficiently and in a way that is appealing and engaging to all stakeholders?
5. What unanticipated positive and negative effects does the instructional intervention produce?
6. Has the instructional environment changed as a result of the intervention?

(Bradley and Reinking, 2011, pp. 314–15)

Similarly, Anderson and Shattuck (2012) note that design-based research is a mixed methods approach which: (a) focuses on interventions in real-world, authentic educational contexts; (b) involve multiple iterations as events evolve; (c) focus on improvements in practice; and (d) seek to test a theory and theoretical relationships (pp. 16–18).

Key principles of design studies in education (The Design-Based Research Collective, 2003, p. 5), for example in developing learning environments, are:

1. They intertwine theory, models and practice.
2. Research and development occur in cycles of refinement, testing and feedback ('design, enactment and analysis'; The Design-Based Research Collective, 2003, p. 6).
3. The findings must be communicated and shared with all parties, including the users.
4. The research and the outcomes must be tested and used in authentic, real-world settings respectively.
5. Reporting and development go together in developing a useable outcome.

Shavelson *et al*. (2003, p. 26) suggest that the key principles of design studies are that they are: (a) 'iterative'; (b) 'process focused'; (c) 'interventionist'; (d) 'collaborative'; (e) 'multileveled'; (f) 'utility oriented'; and (g) 'theory driven'. Cobb *et al*. (2003, p. 9) suggest that theory generation is a key feature of design experiments; they are 'crucibles for the generation and testing of theory' (p. 9), their purpose is to generate theories of teaching and learning (p. 10), and this involves development, intervention and reflection. In having 'pragmatic roots' (p. 10), Cobb *et al*. point us to suggesting the affinity between design experiments and mixed methods research (see Chapter 2; see also Gorard *et al*., 2004, pp. 579, 593). Similarly, Bradley and Reinking (2011) comment that a hallmark of a design experiment is its iterative nature and that this supports the importance accorded to teacher development (p. 307).

The design experiment is perhaps more fittingly termed a 'design study', as it frequently does not conform to the requirements of an experiment (e.g. it does not have the hallmarks of a randomized controlled trial), as set out in the earlier part of this chapter. It is included here because of its nomenclature rather than its affinity to experiments as described in this chapter, though, like an experiment, it involves a deliberate and planned intervention. Anderson and Shattuck (2012) note the increasing interest and growth in design experiments globally, particularly in the US, the Netherlands, the UK and Singapore, and particularly focused on learning interventions, instructional technology and for school-age students.

It is more useful to focus on the word 'design' rather than 'experiment' here, as a design study owes some of its pedigree to engineering and science rather than to an experiment which has control and experimental groups. Brown (1992, p. 141) suggests that design studies attempt to 'engineer innovative educational environments and simultaneously conduct experimental studies of those innovations'. Take the example of engineering: here the designer develops a product and then tests it in real conditions (Gorard *et al*., 2004, p. 576), noting, during the testing (the experiment), what are the problems with the design, what needs to be improved, where there are faults and failures, and so on, gathering data from other participants and users. Then the engineer redesigns the product to address the faults found, refines the product and re-tests the improved product, noting faults, problems or failures; the engineer reworks the product to address these problems and tests it out again, and so on. We can observe here (e.g. Bradley and Reinking, 2011) that:

- the process is iterative; it has many cycles, trials, improvements and refinements over time;
- it focuses on the processes involved in the workings of the product;
- it communicates with different parties (theoreticians, designers, practitioners) about the design and development of the product (the designers, the engineers, the users), akin to a research-and-development model;
- the product has to work in the 'real world' (an example of 'what works') and in non-laboratory conditions and contexts;
- it is data-driven – the next cycle of refinement is based on data (e.g. observational data, measurement data, notes and records) derived from the previous round.

Bradley and Reinking (2011) note that, in addition to the engineering metaphor, the design experiment also emphasizes the ecology metaphor (p. 316), as each classroom has its own ecological character that has to feature in the design experiment. The design experiment has not only to take account of the classroom ecology but must work with it.

The inception of design studies is often attributed to Brown (1992), whose autobiographical account of her years of research charts a movement away from the laboratory and into the classroom, in order to catch the authenticity of the real world in research and development. She recognizes that this is bought at the price of tidiness, and she justifies this in terms of the real world being 'rarely isolatable' in terms of its components, and in which 'the whole really is more than the sum of its parts' (p. 166). For Brown and her successors, interventions are based on theoretical claims (e.g. about teaching and learning) and are inextricably linked to practices that improve the situation (e.g. of teaching and learning); they respond to 'emergent features' of the situation in which they are operating (The Design-Based Research Collective, 2003, p. 6). Practitioners, researchers and developers work together to produce a useful intervention and innovation.

A design-based study focuses on changing practice, instead of the static, 'frozen' input–output model of an intervention that one sees in much experimental and educational research (The Design-Based Research Collective, 2003, p. 7); in a design study the 'product' changes over time, as refinements are made in response to feedback from all parties.

However, unlike an engineering product, a design-based study does not end with the perfecting of a particular product. Rather, as Brown (1992) indicates, it affects theory, for example, of learning, of teaching. The design-based study can address and generate many kinds of knowledge (The Design-Based Research Collective, 2003, p. 8):

- investigating possibilities for new and innovative teaching environments;
- developing theories of teaching and learning that are rooted in real-world contexts;
- developing cumulative knowledge of design;
- increasing capacity in humans for innovation.

To this Shavelson *et al*. (2003, p. 28) suggest that design studies can address research that asks 'what is happening?', 'is there a systematic effect?' and 'why or how is it happening?'

The attraction of the approach is that it takes account of the complex, real, multivariate world of learning, teaching and education; as such, design studies are 'messier' than conventional experiments, as they take account of many variables and contexts, and the intervention develops and changes over time and involves several parties and strives to ensure that what works at the design stage really works in practice (Gorard *et al*., 2004, pp. 578, 582). The design study develops a profile of multiple variables rather than testing a sole hypothesis (Lobato, 2003, p. 19).

On the other hand, Shavelson *et al*. (2003) argue that design studies are not exempted from the usual warrants of research, for example, the epistemological basis and warrants in the research (p. 25), particularly if there are many possible confounding variables at work (p. 27), how generalizable the results can be, as they are so rooted in specific contexts (p. 27), and how alternative explanations of the outcomes have been considered (p. 27). To answer these questions, McCandliss *et al*. (2003, p. 15) also add that video-recording can provide useful data over time, and Shavelson *et al*. (2003) suggest that longitudinal narrative data are particularly useful as they can track developments and causal developments over time in a way that catches the complexity and contextualization of the intervention which inheres in its very principles. However, narrative accounts risk circularity, and there need to be external checks and balances, controls and warrants, in validating the knowledge claims (p. 27).

Further, Sloane and Gorard (2003) and Anderson and Shattuck (2012) indicate some difficulties that design studies have to address, including measurement problems, external validity, the lack of controls and control groups, the problem of insider research (e.g. bias), the lack of failure criteria (and they argue that engineers include failure criteria as essential features of their research and development) and the need for appropriate modelling of causality at both the alpha stages (the designers) and the beta stages (the users). Hence researchers using a design study have to be clear on its purposes, intended contribution to theory generation, participants and communication processes between them, processes of intervention and debriefing/feedback, understanding of the local context of the intervention (Cobb *et al*., 2003, p. 12) and 'testable conjectures' that can be revised iteratively (p. 11).

20.11 Internet-based experiments

A growing field in educational and psychological research is the use of the Internet for experiments. Internet-based experiments adhere to the same principles as 'true' and field experiments, with attention to independent variables, controls and manipulation of the key variable. Hewson *et al*. (2003) and Johnson and Christensen (2010) note that Internet-based experiments have the attractions of: ease of access to diverse and dispersed populations; ease of access by the participants (i.e. they do not have to come to the researcher); high statistical power because of large samples; the opportunity for many participants to be involved simultaneously; access at any time of the day/week; high-speed, real-time access and involvement (and the ability for the researcher to control timing); freedom from experimenter bias (as the researcher is not present); anonymity of the participants; and cost savings.

On the other hand, the researcher has less experimental control; no control over possible multiple, repeated submissions by participants; problems of self-selection (e.g. a non-representative volunteer sample); no control over the experimental conditions and environment in which the involvement takes place; hacking; no control over whether the participants are being honest about themselves and their characteristics; no control over whether the participants are completing the experiment alone or with others; technical problems (e.g. connectivity, compatible software); misunderstandings or lack of understanding of aspects of the experiment by the participants; and dropout. (Indeed it

may be impossible for respondents who withdraw partway through an experiment to have their data withdrawn, as their particular data may not be identifiable (Brooks *et al*., 2014, pp. 72–3). This problem is not exclusive to Internet experiments; it may be the same for other forms of research in which individuals are not required to identify themselves, in the interests of anonymity and non-traceability.) Further, asking young persons to interact with an unknown researcher online may violate the ethical issue of advice given to young people to avoid talking to strangers (p. 94).

Hewson *et al*. (2003, p. 48) classify Internet experiments into four principal types: (i) those using printed materials; (ii) those using non-printed materials such as audio or video; (iii) online reaction-time experiments; and (iv) experiments which require interpersonal interaction.

The first kind of experiment is akin to a survey in that it sends formulated material to respondents (e.g. graphically presented material) by email or by web page, and the intervention will be to send different groups different materials. Here all the cautions and comments that were made about Internet surveys apply (Chapter 18), particularly the problems of download times and different browsers and platforms. However, the matter of download time applies more strongly to the second type of Internet-based experiments that use video clips or sound, and some software packages will reproduce higher quality than others, even though the original that is transmitted is the same for everyone. This can be addressed by ensuring that the material runs at its optimum even on the slowest computer (Hewson *et al*., 2003, p. 49) or by stating the minimum hardware required for the experiment to be run successfully.

Reaction-time experiments, those that require very precise timing (e.g. to milliseconds), are difficult in remote situations, as different platforms and Internet connection speeds and congestion on the Internet through having multiple users at busy times can render standardization virtually impossible. One solution to this is to have the experiment downloaded and then run offline before loading it back onto the computer and sending it.

The fourth type involves interaction, and is akin to Internet interviewing (discussed below), facilitated by chat rooms. However, this is solely a written medium and so intonation, inflection, hesitancies, non-verbal cues, extra-linguistic and paralinguistic factors are ruled out of this medium. It is, in a sense, incomplete, though the use of screen-top video cameras mitigates this. Indeed this latter development renders observational studies an increasing possibility in the Internet age.

Reips (2002a) reports that in comparison to laboratory experiments, Internet-based experiments experienced greater problems of dropout, the dropout rate in an Internet experiment was very varied (from 1 per cent to 87 per cent) and dropout could be reduced by offering incentives, for example, payments or lottery tickets, bringing a difference of as much as 31 per cent to dropout rates. Dropout on Internet-based research was due to a range of factors (e.g. motivation, how interesting the experiment was), not least of which was the non-compulsory nature of the experiment (in contrast, for example, to the compulsory nature of experiments undertaken by university student participants as part of their degree studies). The discussion of the 'high-hurdle technique' (Chapter 18) is applicable to experiments. Reips (2002b, pp. 245–6) also reports that greater variance in results is likely in an Internet-based experiment than in a conventional experiment due to technical matters (e.g. network connection speed, computer speed, multiple software running in parallel). He also reports (Reips, 2009, p. 381) that Internet experiments suffer from reducing the controls that the experimenter can place on the participant and the problems of a biased, volunteer-only sample (p. 382) or recruitment biases.

On the other hand, Reips (2002b, p. 247) also reports that Internet-based experiments have an attraction over laboratory and conventional experiments in that they:

- have greater generalizability because of their wider sampling;
- demonstrate greater ecological validity as typically they are conducted in settings which are familiar to the participants and at times suitable to the participant ('the experiment comes to the participant, not vice versa'), though, of course, the obverse of this is that the researcher has no control over the experimental setting (p. 250);
- they have a high degree of voluntariness, such that more authentic behaviours can be observed.

How correct these claims are is an empirical matter. For example, some software packages can reduce experimenter control as these packages may interact with other programming languages. Indeed Schwarz and Reips (2001) report that the use of Javascript led to a 13 per cent higher dropout rate in an experiment compared to an identical experiment that did not use Javascript. Further, multiple returns by a single participant could confound reliability (see also Chapter 18).

Reips (2002a, 2002b) provides a series of 'dos' and 'don'ts' in Internet experimenting. In terms of 'dos' he gives five main points:

1. Use dropout as a dependent variable.
2. Use dropout to detect motivational confounding (i.e. to identify boredom and motivation levels in experiments).
3. Place questions for personal information at the beginning of the Internet study. Reips (2002b) suggests that asking for personal information may assist in keeping participants in an experiment, and that this is part of the 'high-hurdle' technique, where dropouts self-select out of the study, rather than dropping out during the study.
4. Use techniques that help ensure quality in data collection over the Internet (e.g. the 'high-hurdle' and 'warm-up' techniques discussed earlier, subsampling to detect and ensure consistency of results, using single passwords to ensure data integrity, providing contact information, reducing dropout).
5. Use Internet-based tools and services to develop and announce your study (using commercially produced software to ensure that technical and presentational problems are overcome). Some websites (e.g. the American Psychological Society) also announce experiments.

In terms of 'don'ts' he gives five main points:

1. Do not allow external access to unprotected directories. This can violate ethical and legal requirements, as it provides access to confidential data. It also might allow the participants to have access to the structure of the experiment, thereby contaminating the experiment.
2. Do not allow public display of confidential participant data through URLs (a problem as these can be found easily), as this again violates ethical codes.
3. Do not accidentally reveal the experiment's structure (as this could affect participant behaviour). This might be done through including the experiment's details on a related file or a file in the same directory.
4. Do not ignore the technical variance inherent in the Internet (configuration details, browsers, platforms, bandwidth and software might all distort the experiment, as discussed above).
5. Do not bias results through improper use of form elements (i.e. measurement errors, where omitting particular categories (e.g. 'neutral', 'do not want to respond', 'neither agree nor disagree') could distort the results).

The points made in connection with Internet surveys and questionnaires (Chapters 18 and 24) apply equally to Internet experiments, and we advise readers to review these.

Reips (2002b) points out that it is a misconception to regard an Internet-based experiment as the same as a laboratory experiment, as: (a) Internet participants can choose to leave the experiment at any time; (b) they can conduct the experiment at any time and in their own settings; (c) they are often conducted with larger samples than conventional experiments; (d) they rely on technical matters, network connections and the computer competence of the participants; and (e) they are more public than most conventional experiments. On the other hand, he also cautions against regarding the Internet-based experiment as completely different from the laboratory experiment, as: (a) many laboratory experiments also rely on computers; (b) fundamental ideas are the same for laboratory and Internet-based surveys; (c) similar results have been produced by both means. He suggests several issues in conducting Internet-based experiments:

- consider a web-based software tool to develop the experimental materials;
- pilot the experiment on different platforms for clarity of instructions and availability on different platforms;
- decide the level of sophistication of HMTL scripting and whether to use HTML or non-HTML;
- check the experiments for configuration errors and variance on different computers;
- place the experiment on several websites and services;
- run the experiment online and offline to make comparisons;
- use the 'warm-up' and 'high-hurdle' techniques, asking filter questions (e.g. about the seriousness of the participant, their background and expertise, language skills);
- use dropout to ascertain whether there is motivational confounding;
- check for obvious naming of files and conditions (to reduce the possibility of unwanted access to files);
- consider using passwords and procedures (e.g. consistency checks) to reduce the possibility of multiple submissions;
- keep an experimental log of data for any subsequent analysis and verification of results;
- analyse and report dropout;
- keep the experimental details on the Internet, to give a positive impression of the experiment.

Reips (2009, p. 375) also writes that the success of Internet-based experimentation depends in part on the 'cues transmitted', the 'bandwidth', 'cost constraints', 'level and type of anonymity' and 'synchronicity and exclusivity'.

Given the rise of evidence-based practice in education, and the advocacy of randomized controlled trials in education, this form of experimentation has become more widely used in education. We also refer readers to Birnbaum (2009) and Joinson *et al*. (2009).

20.12 *Ex post facto* research

Ex post facto studies start with groups that are already different with regard to certain characteristics or observations; here the researcher goes in reverse, searching back for likely factors that brought about those differences.

In *ex post facto* experiments, it is not possible to control variables in advance of the experiment or during the experiment, the data being already in existence before the experiment has commenced. However, in this case, the controls can be applied at the stage of data analysis, where the researcher can manipulate the independent variables to hold them constant, i.e. to control for the relative effects of these. For an example of this, we refer the reader to Chapter 6 on causation, and to Chapter 40 for an indication on how controls can be placed statistically, for example, partial correlations and crosstabulations.

In introducing *ex post facto* research here, we focus on its key features and how to conduct such a project, including:

- co-relational and criterion groups designs;
- characteristics of *ex post facto* research;
- occasions when appropriate;
- advantages and disadvantages of *ex post facto* research;
- designing an *ex post facto* investigation;
- procedures in *ex post facto* research.

In *ex post facto* research the researcher takes the effect (or dependent variable) and examines the data retrospectively to establish causes, relationships or associations, and their meanings.

Introduction

When translated literally, *ex post facto* means 'after the fact'; it signifies 'from what is done afterwards', 'from after the event' or 'from what has happened'. In the context of social and educational research, the phrase means 'retrospectively' and refers to those studies which investigate possible cause-and-effect relationships by observing an existing condition or state of affairs and searching back in time for plausible causal factors. In terms of Chapter 6 (on causation), this is examining the causes of effects, and we advise readers to refer to that chapter. Here researchers ask themselves what factors seem to be associated with certain occurrences, conditions or aspects of behaviour. As they have happened already, the researcher has to hypothesize possible causes and then test them against the evidence, for example, by holding factors constant and by controlling and matching the samples.

Ex post facto research is a method of teasing out possible antecedents of events that have happened and cannot, therefore, be controlled, engineered or manipulated by the investigator (Cooper and Schindler, 2001, p. 136). Researchers can only report what has happened or what is happening, by trying to hold factors constant by careful attention to sampling. Independent variables cannot be manipulated as in true experiments, as they have already happened. Hence the researcher is in the realms of probabilistic causation, inferring causes tentatively rather than being able to demonstrate causality unequivocally.

Ex post facto research can be used to study groups which are similar and which have had the same experience with the exception of one condition, and here the effect of the one differing condition on the dependent variable can be assessed. *Ex post facto* research, then, is a form of experiment, but without the stringent controls of a true experiment; there are control and experimental groups (the latter where a particular condition has been applied), but, since there is little or no rigorous manipulation of the independent variables or conditions, and since there is no random allocation of subjects to groups, any inferences of causation are tentative.

The following example will illustrate the basic idea. Let us return to the example introduced earlier in this chapter. Imagine a situation in which there has been a dramatic increase in the number of fatal road accidents in a particular locality. An expert is called in to investigate. Naturally, there is no way in which she can set the actual accidents because they have already happened; nor can she turn to technology for a video replay of the incidents; nor can she require a participant to run under a bus or a lorry, or to stand in the way of a speeding bicycle, in order to discover the effects. What she can do, however, is to study hospital records to see which groups have experienced the greatest trauma – bus, lorry or bicycle impact victims. Or she can attempt a reconstruction by studying the statistics, examining the accident spots and taking note of the statements given by victims and witnesses. In this way the expert will be in a position to identify possible determinants of the accidents, looking at the outcomes and working backwards to examine possible causes. These may include excessive speed, poor road conditions, careless driving,

frustration, inefficient vehicles, effects of drugs or alcohol and so on. On the basis of her examination, she can formulate hypotheses as to the likely causes and submit them to the appropriate authority in the form of recommendations. These may include improving road conditions, or lowering the speed limit, or increasing police surveillance, for instance. The point of interest to us is that in identifying the causes retrospectively, the expert adopts an *ex post facto* perspective.

Ex post facto research is a method that can also be used instead of an experiment, to test hypotheses about cause and effect in situations where it is impossible, impractical or unethical to control or manipulate the dependent variable or, indeed, the independent variables. For example, let us say that we wish to test the hypothesis that family violence causes poor school performance. Here, ethically speaking, we should not expose a student to family violence. However, one could put students into two groups, matched carefully on a range of factors, with one group comprising those who have experienced family violence and the other comprising those who have not. If the hypothesis is supportable then the researcher should be able to discover a difference in school performance between the two groups when the other variables are matched or held as constant as possible.

Kerlinger (1970) has defined *ex post facto* research as that in which the independent variable or variables have already occurred and in which the researcher starts with the observation of a dependent variable or variables. She then studies the independent variable or variables in retrospect for their possible relationship to, and effects on, the dependent variable or variables. The researcher is thus examining retrospectively the effects of a naturally occurring event on a subsequent outcome with a view to establishing a causal link between them. The key to establishing the causes is the careful identification of those that are possible, testing each against the evidence, and then eliminating the ones that do not stand up to the test, ensuring that attention is paid to careful sampling and to controls – holding fixed some variables.

Some instances of *ex post facto* designs correspond to experimental research in reverse, for instead of taking groups that are equivalent and subjecting them to different treatments so as to bring about differences in the dependent variables to be measured, an *ex post facto* experiment begins with groups that are already different in some respect and searches in retrospect for the factor that brought about the difference. An *ex post facto* experiment, then, is a form of quasi-experiment.

One can discern two approaches to *ex post facto* research. In the first approach one commences with subjects who differ on an *independent* variable, for example, their years of study in mathematics, and then studies how they differ on the dependent variable, for example, a mathematics test. In a second approach, one can commence with subjects who differ on the *dependent* variable (e.g. their performance in a mathematics test) and discover how they differ on a range of independent variables, for example, their years of study, their liking for the subject, the amount of homework they do in mathematics. The *ex post facto* research here seeks to discover the causes of a particular outcome (mathematics test performance) by comparing those students in whom the outcome is high (high marks on the mathematics test) with students whose outcome is low (low marks on the mathematics test), after the independent variable has occurred.

Ary *et al*. (2006, p. 335) discuss 'proactive' and 'retroactive' *ex post facto* research designs. In the former, the subjects are grouped on the basis of the presence or absence of an independent variable, and then the researcher compares the groups in terms of the outcomes – the dependent variable. In the latter, the dependent variable is constant, and the researcher seeks to discover the independent variables that might have contributed to the outcome, hypothesizing about these independent variables and then testing them against the evidence. Figure 20.6 indicates these two main types of *ex post facto research* designs.

Here is an example of an *ex post facto* piece of research. It has been observed that staff at a very large secondary school have been absent on days when they teach difficult classes. An *ex post facto* piece of research was conducted to try to establish the causes of this. Staff absences on days when teaching difficult secondary classes were noted, thus:

	Days when teaching difficult secondary classes	
Absences	Yes	No
High	26	30
Low	22	50
Total	48	80
Overall total: 128		

Here the question of time was important: were the staff absent only on days when they were teaching difficult classes or at other times? Were there other variables that could be factored into the study, for example, age groups? Hence the study was refined further, collecting more data:

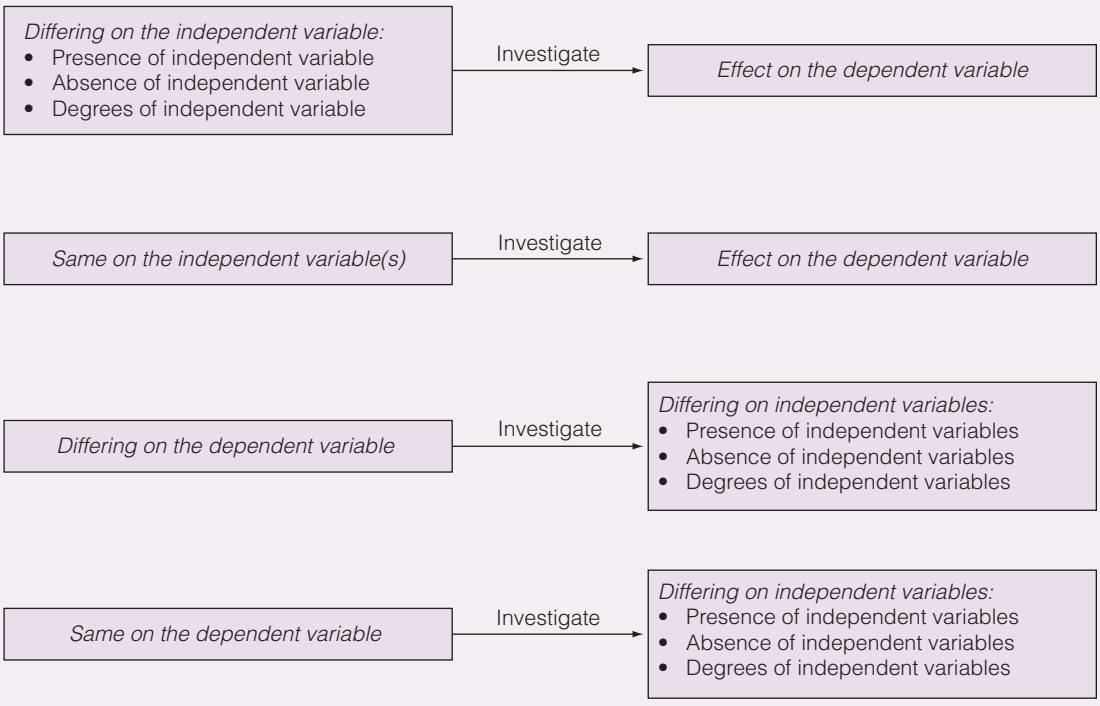


FIGURE 20.6 Four types of *ex post facto* research

Age	*Days when teaching difficult secondary classes*		*Days when not teaching difficult secondary classes*	
	High absence	*Low absence*	*High absence*	*Low absence*
<30 years old	30	6	16	10
30–50 years old	4	4	4	20
>50 years old	2	2	2	28
Total	36	12	22	58
Overall total: 128				

This shows that age was also a factor as well as days when teaching difficult secondary classes: younger people were more likely to be absent. Most teachers who were absent were under thirty years of age. Within age groups, it is also clear that young teachers had a higher incidence of excessive absence when teaching difficult secondary classes than teachers of the same (young) age group when they were not teaching difficult secondary classes.

Of course, a further check here would be to compare the absence rates of the same teachers when they did and did not teach difficult classes, and conduct difference tests (e.g. t-tests, ANOVA: see Chapter 41) to examine differences between the two sets of scores (days when difficult classes were taught and days when they were not taught; differences between age groups in respect of the days when difficult classes were and were not taught).

Co-relational and criterion groups designs

Two kinds of design may be identified in *ex post facto* research – the co-relational study and the criterion group study. The former is sometimes termed 'causal research' and the latter, 'causal-comparative research'. A co-relational (or causal) study is concerned with identifying the antecedents of a present condition. As its name suggests, it involves the collection of two sets of data, one of which will be retrospective, with a view to determining the relationship between them. The basic design of such an experiment can be represented thus (using the symbols from Campbell and Stanley (1963), where X = the independent variable and O = the dependent variable, discussed below):

$$X \rightarrow O$$

A study by Borkowsky (l970) was based upon this kind of design. He attempted to show a relationship between the quality of a music teacher's undergraduate training (*X*) and his subsequent effectiveness as a teacher of his subject (*O*). Measures of the quality of a music teacher's college training included grades in specific courses, overall grade average and self-ratings, etc. Teacher effectiveness was assessed by indices of pupil performance, pupil knowledge, pupil attitudes and judgement of experts, etc. Correlations between all measures were obtained to determine the relationship. At most, this study could show that a relationship existed, after the fact, between the quality of teacher preparation and subsequent teacher effectiveness. Where a strong relationship is found between the independent and dependent variables, three possible interpretations are open to the researcher:

1. that the variable *X* has caused *O*;
2. that the variable *O* has caused *X*; or
3. that some third unidentified, and therefore unmeasured, variable has caused *X* and *O*.

It is often the case that a researcher cannot tell which of these is correct. This raises the issue of the direction of causality: it is difficult in an *ex post facto* experiment to determine what causes what: whether A causes B or B causes A.

The value of co-relational or causal studies lies chiefly in their exploratory or suggestive character, for while they are not always adequate in themselves for establishing causal relationships among variables, they are a useful first step in this direction in that they do yield measures of association.

In the criterion-group (or causal-comparative) approach, the investigator sets out to discover possible causes of a phenomenon being studied, by comparing the subjects in which the variable is present with similar subjects in whom it is absent, i.e. noting the circumstances in which a given effect occurs and does not occur (Lord, 1973, p. 3). The basic design in this kind of study may be represented thus:

	O_1
X	
	O_2

If, for example, a researcher chose such a design to investigate factors contributing to teacher effectiveness, the criterion group O_1, the effective teachers, and its counterpart O_2, a group *not* showing the characteristics of the criterion group, are identified by measuring the differential effects of the groups on classes of children. The researcher may then examine *X*, some variable or event, such as the background, training, skills and personality of the groups, to discover what might 'cause' only some teachers to be effective.

Morrison (2009, p. 181) gives an example of a criterion-group piece of *ex post facto* research. He writes thus:

> Let us imagine, for example, that the researcher is seeking to establish the cause of effective teaching, and hypothesizes that one cause is collegial curriculum planning with other members of the department. The research could be designed as in Figure 20.7.
>
> Here there are two criterion groups: (a) the presence of collegial curriculum planning; and (b) the absence of collegial curriculum planning. By examining the difference in teaching effectiveness between those teachers (however one wished to measure 'effective teaching') who did and did not plan their curriculum with colleagues (collegial curriculum planning) one could infer a possible causal difference. But one has to be cautious: at most this is a correlational study and causation is not the same as correlation. Indeed … a third cause may be influencing both the effective/ineffective teaching and the presence/absence of collegial curriculum planning, e.g. staff sociability.
>
> (Morrison, 2009, p. 181)

The causal-comparative design is different from a historical design, in that the former is concerned with present events, whereas the latter traces the history of past events (Lord, 1973, p. 4).

Criterion-group or causal-comparative studies may be seen as bridging the gap between descriptive research methods on the one hand and true experimental research on the other.

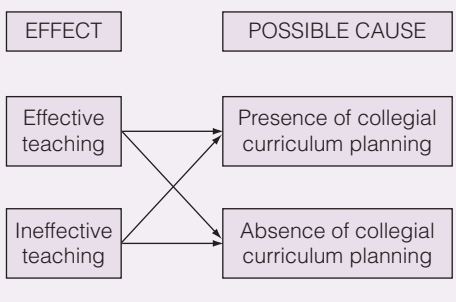


FIGURE 20.7 Two causes and two effects

Controls and causality

Other characteristics of *ex post facto* research become apparent when it is contrasted with true experimental research. Kerlinger (1970) describes the modus operandi of the experimental researcher. ('If *x*, then *y*' in Kerlinger's usage. We have substituted *X* for *x* and *O* for *y* to fit in with Campbell and Stanley's (1963) conventions throughout the chapter.) Kerlinger hypothesizes: if *X*, then *O*; if frustration, then aggression. Depending on circumstances and his own predilections in research design, he uses some method to manipulate *X*. He then observes *O* to see if concomitant variation, the variation expected or predicted from the variation in *X*, occurs (see also Chapter 6). If it does, this is evidence for the validity of the proposition 'if *X*, then *O*'. Note that the scientist here predicts from a controlled *X* to *O*. To help him achieve control, he can use the principle of randomization and active manipulation of *X* and can assume, other things being equal, that *O* is varying as a result of the manipulation of *X*.

In *ex post facto* designs, on the other hand, *O* is observed. Then a retrospective search for *X* ensues. An *X* is found that is plausible and agrees with the hypothesis. Due to lack of control of *X* and other possible *X*s, the truth of the hypothesized relation between *X* and *O* cannot be asserted with the confidence of the experimental researcher. Basically, then, *ex post facto* investigations have, so to speak, a built-in weakness: lack of control of the independent variable or variables. As Spector (1993, p. 43) suggests, it is impossible to isolate and control every possible variable, or to know with absolute certainty which are the most crucial variables.

The most important difference between experimental and *ex post facto* designs is control. In the experimental situation, investigators at least have manipulative control; they have as a minimum one active variable. If an experiment is a 'true' experiment, they can also exercise control by randomization. They can assign subjects to groups randomly; or, at the very least, they can assign treatments to groups at random. In the *ex post facto* research situation, this control of the independent variable is not possible, and, perhaps more important, neither is randomization. Investigators must take things as they are and try to disentangle them, though having said this, they can make use of selected procedures that will give them an element of control in this research. These we shall touch upon shortly.

By their very nature, *ex post facto* experiments can provide support for any number of different, perhaps even contradictory, hypotheses; they are so flexible that it is largely a matter of postulating hypotheses according to one's personal preference. The investigator begins with certain data and looks for an interpretation consistent with them; often, however, a number of interpretations may be at hand. Consider again the hypothetical increase in road accidents in a given town. A retrospective search for causes will disclose half-a-dozen plausible ones.

Experimental studies, by contrast, begin with a specific interpretation and then determine whether it is congruent with externally derived data. Frequently, causal relationships seem to be established on nothing more substantial than the premise that any related event occurring prior to the phenomenon under study is assumed to be its cause – the classical *post hoc, ergo propter hoc* fallacy ('after this, therefore because of this'); just because one variable precedes another in time, it does not follow that the first variable *causes* the second: I may drink coffee and then have a sleepless night, but it does not follow that drinking the coffee *caused* the sleepless night – there may have been other causes. Even when we do find a relationship between two variables, we must recognize the possibility that both are individual results of a common third factor rather than the first being necessarily the cause of the second.

As mentioned earlier, there is also the real possibility of reverse causation, for example, that a heart condition promotes obesity rather than the other way around, or that they encourage each other. The point is that the evidence simply *illustrates* the hypothesis; it does not test it, since hypotheses cannot be tested on the same data from which they were derived. The relationship noted may actually exist, but it is not necessarily the only relationship, or perhaps the crucial one. Before we can accept that smoking is the primary cause of lung cancer, we have to rule out alternative hypotheses.

Further, a researcher may find that playing computer games correlates with poor school performance. Now, it may be there is a causal effect here: playing computer games causes poor school performance; or there may be reverse causality: poor school performance causes students to playing computer games. However, there may be a third explanation: students who, for whatever reason (e.g. ability, motivation), do not do well at school also like playing computer games; it may be the third variable (the independent variable of ability or motivation) that is causing the other two outcomes (playing computer games or poor school performance).

We cannot conclude from what has just been said that *ex post facto* studies are of little value; many important investigations in education and psychology

are *ex post facto* designs. There is often no choice in the matter: an investigator cannot cause one group to become failures, delinquent, suicidal, brain-damaged or dropouts. Research must of necessity rely on existing groups. On the other hand, the inability of *ex post facto* designs to incorporate the basic need for control (e.g. through manipulation or randomization) makes them vulnerable from a scientific point of view and the possibility of their being misleading should be clearly acknowledged. Indeed, *ex post facto* designs are probably better conceived more circumspectly, not as experiments with the greater certainty that these denote, but more as surveys, useful as sources of hypotheses to be tested by more conventional experimental means at a later date.

Occasions when appropriate

Ex post facto designs are appropriate in circumstances where the more powerful experimental method is not possible. These arise when, for example, it is not possible to select, control and manipulate the factors necessary to study cause-and-effect relationships directly; or when the control of all variables except a single independent variable may be unrealistic and artificial, preventing the normal interaction with other influential variables; or when laboratory controls for many research purposes would be impractical, costly or ethically undesirable.

Ex post facto research is particularly suitable in social, educational and psychological contexts where the independent variable or variables lie outside the researcher's control. Examples of the method abound here: the research on cigarette-smoking and lung cancer, for instance; or studies of teacher characteristics; or studies examining the relationship between political and religious affiliation and attitudes; or investigations into the relationship between school achievement and independent variables such as social class, ethnicity, gender and intelligence. Such investigations may be large scale or small scale *ex post facto*.

For educational research, public domain databases and data sets can be used for conducting *ex post facto* educational research, for example, databases and data sets produced by:

- government agencies (e.g. www.gov.uk/government/statistics);
- research agencies (e.g. www.data-archive.ac.uk);
- consortia (e.g. www.socsciresearch.com/r6.html);
- organizations, for example:
  - The OECD: http://stats.oecd.org/index.aspx;
  - UNESCO (Institute for Statistics);
  - The PISA database (www.oecd.org/pisa/pisaproducts; https://nces.ed.gov/surveys/pisa);
  - The World Bank (statistics section);
  - The TIMSS database (http://nces.ed.gov/timss/datafiles.asp).

Advantages and disadvantages of *ex post facto* research

Among the advantages of the approach are the following:

- *ex post facto* research meets an important need of the researcher where the more rigorous experimental approach is not possible;
- the method yields useful information concerning the nature of phenomena – what goes with what and under what conditions. Here *ex post facto* research is a valuable exploratory tool;
- improvements in statistical techniques and general methodology have made *ex post facto* designs more defensible;
- in some ways and in certain situations the method is more useful than the experimental method, especially where the setting up of the latter would introduce a note of artificiality into research proceedings;
- *ex post facto* research is particularly appropriate when simple cause-and-effect relationships are being explored;
- the method can give a sense of direction and provide a fruitful source of hypotheses that can subsequently be tested by the more rigorous experimental method.

Among the limitations and weaknesses of *ex post facto* designs are the following:

- there is the problem of lack of control in that the researcher is unable to manipulate the independent variable or to randomize her subjects;
- one cannot know for certain whether the causative factor has been included or even identified;
- it may be that no single factor is the cause;
- a particular outcome may result from different causes on different occasions;
- when a relationship has been discovered, there is the problem of deciding which is the cause and which the effect; the possibility of reverse causation must be considered;
- the relationship of two factors does not establish cause and effect;
- the *ex post facto* hypothesis is generated after the data have been collected, so it is not possible to disconfirm it (Babbie, 2010, p. 462);

- classifying into dichotomous groups can be problematic;
- there is the difficulty of interpretation and the danger of the *post hoc* assumption being made, that is, believing that because *X* precedes *O*, *X* causes *O*;
- as the researcher attempts to match groups on key variables, this leads to shrinkage of sample (Spector, 1993, p. 43). (Lewis-Beck (1993, p. 43) reports an example of such shrinkage from a sample of 1,194 to 46 after matching had been undertaken);
- it often bases its conclusions on too limited a sample or number of occurrences;
- it frequently fails to single out the really significant factor or factors, and fails to recognize that events have multiple rather than single causes;
- as a method it is regarded by some as too flexible;
- it lacks nullifiability and confirmation.

Designing an *ex post facto* investigation

We earlier referred to the two basic designs embraced by *ex post facto* research – the co-relational (or causal) model and the criterion group (or causal-comparative) model. As we saw, the causal model attempts to identify the antecedent of a present condition and may be represented thus:

Independent variable	*Dependent variable*
X	O

Although one variable in an *ex post facto* study cannot be confidently said to depend upon the other as would be the case in a truly experimental investigation, it is nevertheless usual to designate one of the variables as independent (*X*) and the other as dependent (*O*). The left to right dimension indicates the temporal order, though having established this, we must not overlook the possibility of reverse causality. In a typical investigation of this kind, then, two sets of data relating to the independent and dependent variables respectively are gathered. As indicated earlier, the data on the independent variable (*X*) will be retrospective in character and as such will be prone to the kinds of weakness, limitations and distortions to which all historical evidence is subject.

For example, imagine a secondary school in which it is hypothesized that low staff morale (*O*) has come about as a direct result of school reorganization some two years earlier. A number of key factors distinguishing the new organization from the previous one can be identified. Collectively these could represent or contain the independent variable *X* and data on them could be accumulated retrospectively, for example, curricular innovation, loss of teacher status, decline in student motivation, modifications to the school catchment area or the appointment of a new headteacher. These could then be checked against a measure of prevailing teachers' attitudes (*O*), thus providing the researcher with some leads at least as to possible causes of current discontent.

Here the causal-comparative model may be represented schematically as:

Group	*Independent variable*	*Dependent variable*
E	X	O_1
C		O_2

Using this model, the investigator hypothesizes the independent variable and then compares two groups, an experimental group (*E*) which has been exposed to the presumed independent variable *X* and a control group (*C*) which has not. (The dashed line in the model shows that the comparison groups *E* and *C* are not equated by random assignment.) Alternatively, she may examine two groups that are different in some way or ways and then try to account for the difference or differences by investigating possible antecedents. We refer the reader to Chapter 6 on effect-to-cause investigations.

The basic design of causal-comparative investigations is similar to an experimentally designed study. The chief difference resides in the nature of the independent variable, *X*. In a truly experimental situation, this will be under the control of the investigator and may therefore be described as manipulable. In the causal-comparative model (and also the causal model), however, the independent variable is beyond her control, having already occurred. It may therefore be described in this design as non-manipulable.

Procedures in *ex post facto* research

Ex post facto research is concerned with discovering relationships among variables in one's data; and we have seen how this may be accomplished by using either a causal or causal-comparative model. We now examine the steps involved in implementing a piece of *ex post facto* research. We begin by identifying the problem area to be investigated. This stage will be followed by a clear and precise statement of the hypothesis to be tested or questions to be answered. The next step is to make explicit the assumptions on which the hypothesis and subsequent procedures will be based. A review of the research literature follows. This enables the investigator to ascertain the kinds of issues, problems, obstacles and findings disclosed by previous

studies in the area. There then follows the planning of the actual investigation and this consists of three broad stages – identification of the population and samples; the selection and construction of techniques for collecting data; and the establishment of categories for classifying the data. The final stage involves the description, analysis and interpretation of the findings.

Drawing on Lord (1973, p. 6), we can set out several stages in conducting an *ex post facto* piece of research:

Stage 1: Define the problem and survey the literature.
Stage 2: State the hypotheses and the assumptions or premises on which the hypotheses and research procedures are based.
Stage 3: Select the subjects (sampling) and identify the methods for collecting the data.
Stage 4: Identify the criteria and categories for classifying the data to fit the purposes of the study and which are as unambiguous as possible and which will enable relationships and similarities to be found.
Stage 5: Gather data on those factors which are always present in which the given outcome occurs, and discard the data in which those factors are not always present.
Stage 6: Gather data on those factors which are always present in which the given outcome does not occur.
Stage 7: Compare the two sets of data (i.e. subtract the former (Stage 5) from the latter (Stage 6)), in order to be able to infer the causes that are responsible for the occurrence or non-occurrence of the outcome.
Stage 8: Analyse, interpret and report the findings.

One has to bear in mind that the evidence illustrates rather than tests the hypothesis here (Lord, 1973, p. 7). It was noted earlier that the principal weakness of *ex post facto* research is the absence of control over the independent variable influencing the dependent variable in the case of causal designs or affecting observed differences between dependent variables in the case of causal-comparative designs. Although the *ex post facto* researcher is denied not only this kind of control but also the principle of randomization, she can nevertheless utilize procedures that provide some measure of control in her investigation; it is to some of these that we now turn.

One of the commonest means of introducing control into this type of research is that of matching the subjects in the experimental and control groups where the design is causal-comparative. Matched pair designs are careful to match the participants on important and relevant characteristics that may have a bearing on the research (for an example of this, see Leow, 2009). There are difficulties with this procedure, however, for it assumes that the investigator knows what the relevant factors are, that is, the factors that may be related to the dependent variable. Further, there is the possibility of losing those subjects who cannot be matched, thus reducing one's sample.

As an alternative procedure for introducing a degree of control into *ex post facto* research, the researcher can build the extraneous independent variables into the design and then use an analysis of variance technique. For example, if intelligence is a relevant extraneous variable but it is not possible to control it through matching or other means, then it could be added to the research as another independent variable, with the participants being classified in terms of intelligence levels. Through analysis of variance techniques the dependent variable measures would then be analysed to reveal the main and interaction effects of intelligence, indicating any statistically significant differences or effect sizes between the groups on the dependent variable, even though no causal relationship between intelligence and the dependent variable could be assumed.

Yet another procedure which may be adopted for introducing a measure of control into *ex post facto* design is that of selecting samples that are as homogeneous as possible on a given variable. For example, if intelligence were a relevant extraneous variable, its effects could be controlled by including participants from only one intelligence level. This would disentangle the independent variable from other variables with which it is commonly associated, so that any effects found could be associated justifiably with the independent variable.

Finally, control may be introduced into an *ex post facto* investigation by stating and testing any alternative hypotheses that might be plausible explanations for the empirical outcomes of the study. A researcher has to beware of accepting the first likely explanation of relationships in an *ex post facto* study as necessarily the only or final one. A well-known instance to which reference has already been made is the presumed relationship between cigarette smoking and lung cancer. Health officials were quick to seize on the explanation that smoking causes lung cancer. Tobacco firms, however, put forward an alternative hypothesis – that both smoking and lung cancer were possibly the result of a third factor, i.e. the possibility that both the independent and dependent variables were simply two separate results of a single common cause.

20.13 Conclusion

This chapter has introduced a range of different types of experiment. Starting with the randomized controlled trial, the 'true' experiment', it held this up as the

clearest example of a full experiment, as it abides by all the features of an experiment that is intended to yield evidence of 'what works'. The strengths and limitations of the true experiment and the randomized controlled trial were set out. Further variants of a true experiment were set out. Moving further out of the laboratory and into the 'real world', the chapter then presented a discussion of different types of quasi-experiment, i.e. those kinds of experiment in which not all the requirements of a true experiment were met or, in the case of *ex post facto* experiments or those which could not be justified on ethical or practical grounds, where the requirements of a true experiment were impossible to meet. Rendering an experiment a quasi-experiment rather than a true experiment was seen to lie not only in design matters, but also in issues of sampling and controls. The chapter introduced design experiments, or, as was argued to be more fittingly described, a design study, Internet experiments and their limitations, and a full overview of *ex post facto research*.

Notes

1 Maynard and Chalmers (1997); Brown *et al*.'s (2011) study of bullying prevention; Slavin *et al*.'s (2009) study of a middle school cooperative reading programme; Tracey *et al*.'s (2010) study of cooperative learning's effects on students' mathematics achievement; Madden *et al*.'s (2011) study of cooperative writing; Buckingham *et al*.'s (2012) study of a small group intervention for older low-progress readers; Jennings's *et al*.'s (2013) study of cultivating awareness and resilience in education; the list is endless.

2 Morrison (2001); Shadish *et al*. (2002); Maxwell (2004); Hammersley (2005, 2008, 2015b); Biesta (2007, 2010b); Frueh (2009); Hutchison and Styles (2010); Sullivan (2011); Bouguen and Gurgand (2012); Cartwright and Hardie (2012); Goldacre (2013); Camburn *et al*. (2015).

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Meta-analysis, systematic reviews and research syntheses

CHAPTER 21

Harsh Suri

This chapter will explore:

- meta-analysis and the different stages of research synthesis
- systematic reviews
- methodologically inclusive research syntheses

21.1 Introduction

In contemporary educational research, most issues or interventions tend to be examined in a variety of contexts utilizing a diverse range of methodological approaches. Making sense of such complex domains of literature to inform policy, practice or further research can be challenging for decision makers and practitioners. As evidence-based education gathers pace, research syntheses are increasingly gaining prominence as valid methods for knowledge generation in their own right (Suri, 2014). Most high-ranking educational research journals have become open to publishing quality research syntheses. Many educational research journals also specifically focus on quality research syntheses. These include *Review of Educational Research*, *Australian Education Review*, *Educational Research Review*, *Research Synthesis Methods* and *Review of Education*.

Glass (1976) coined the term 'meta-analysis' to distinguish between three forms of analysis: primary analysis, secondary analysis and meta-analysis (p. 3). Primary research involves collecting and analysing fresh data; secondary analysis involves re-analysing data collected for primary research to answer different questions; and meta-analysis involves rigorous statistical integration of findings reported across a number of primary research studies. Meta-analysts employ explicit protocols to enhance consistency and objectivity through all stages. In the last four decades, popularity of meta-analyses has grown exponentially (Glass, 2006; Cooper and Hedges, 2009).

In popular areas of research, such as the effect of different interventions on student achievement, Hattie *et al*. (2014) note that on average, one primary research report gets published every hour. A number of meta-analyses have been conducted to examine the impact of individual interventions on student achievement. Making rigorous comparisons across the findings of a number of meta-analyses is a complex endeavour for which useful guidelines have been provided (Hattie *et al*., 2014). Hattie's visible learning series, based on the synthesis of more than 800 meta-analyses (Hattie, 2009), has generated significant interest among teachers and policy makers (see visible-learning.org).

While meta-analyses facilitate comparisons across quantitative studies by bringing them on a common metric called effect size, they are not suitable for synthesizing qualitative research. In a seminal research monograph, Noblit and Hare (1988) proposed meta-ethnography as an appropriate method for synthesizing a small number of qualitative research reports selected through the logic of purposeful sampling. Distinguishing features of their approach include an emphasis on being 'interpretive rather than aggregative' (p. 11); being inductive rather than using an a priori conceptual framework; employing purposeful sampling rather than exhaustive sampling for selecting primary research studies; being consciously aware of one's own subjectivity; and paying attention to the target audience's discourse (Noblit and Hare, 1988). The literature on guidelines for rigorous synthesis of qualitative research in education was relatively sparse until the early 2000s. In this century, excellent guidelines have been published by educational researchers for publishing qualitative research (e.g. Major and Savin-Baden, 2010; Suri, 2014).

Following the popularity of systematic reviews of research in medicine, evidence-based education and systematic reviews also became popular in education in the last two decades. In the past two decades, several large-scale centres have been established to support systematic reviews (e.g. Campbell Collaboration, n.d.; EPPI-Centre, n.d.). Early proponents of systematic reviews in education were meta-analysts who used randomized controlled trials as the gold standard for rigour. In recent years a number of systematic reviewers have included qualitative and mixed methods research.

Educational researchers espousing critical and interpretive orientations have made another important contribution to methods of research synthesis by questioning the rhetoric of systematic reviews and aggregative reviews and raising a concern over this rhetoric which undermines the contributions of other forms of conceptual reviews (Hammersley, 2001; Clegg, 2005; Kennedy, 2007).

Research synthesis is an umbrella term which includes a range of styles of bringing together into a single expert review or report several studies and summaries on a particular topic. The evidence, methodological perspectives and techniques employed in a research synthesis can be qualitative, quantitative or a combination of both. The purpose of a research synthesis is to produce new knowledge by making explicit connections and relations between individual reports that were not visible before. It involves purposeful selection, review, analysis and synthesis of previously published reports on a similar topic to draw conclusions that enable recommendations to be made for policy, practice and further research. In a rigorous synthesis, readers are provided with sufficient information about the synthesis process so that they can make informed decisions about the extent to which the synthesis findings may be adapted to their own contexts (Cooper and Hedges, 2009; Suri, 2014).

This chapter introduces key issues in the fields of meta-analysis, systematic reviews and methodologically inclusive research syntheses as part of the push towards evidence-based policy and practice in education.

21.2 Meta-analysis

Early meta-analysts criticized intuitive narrative reviews for not being comprehensive in their coverage; overly relying on significance tests and subjective judgements; being prone to Type II error and inconclusive findings; overlooking the magnitude of the effect sizes; and overlooking contextual variables that potentially moderate effect size (Jackson, 1980; Cook *et al*., 1992).

To reduce unstated subjectivity in aggregating findings from a range of separate and disparate primary research reports examining a similar concept or intervention, meta-analysts recommend an explicit adherence to scientific rigour in each of the following stages of research synthesis (Lipsey and Wilson, 2001; Glass, 2006; Cooper and Hedges, 2009):

- formulating a problem;
- searching for relevant literature;
- extracting relevant information from selected studies;
- integrating findings across studies;
- presenting the findings as a scientific report.

Meta-analysts have systematized the entire process of research synthesis by identifying the main tasks in each phase, highlighting critical decision points within each phase and allowing discussion of the relative merits of different choices at each decision point. They advocate explicit statement and justification of the decisions made at each stage of the research synthesis from hypothesis formulation, data selection, evaluation, analysis and interpretation, to public presentation. There are also several types of sensitivity analyses that can examine the dependence of the findings on the assumptions made about the nature of the data. Over the past three decades, meta-analysts have conducted numerous investigations to examine the robustness of their techniques and have explored ways of refining these techniques, as well as examining many substantive uses in the field of education. Meta-analysis now has become but one (important) method in integrative research synthesis.

Formulating a problem

Meta-analyses seek to discover both consistencies in similar-appearing primary studies and also to account for the variability found between them, leading to generalizations within the limits and contexts of the research studies used (Cooper and Hedges, 2009). The purpose of a meta-analysis is formulated in terms of a clear hypothesis with conceptual and operational definitions of key constructs, independent variables and dependent variables. Keeping in mind the intended audience, the contextual variation covered within the scope of the synthesis is explicitly stated. Meta-analysts note that statistical significance is easier to achieve with large samples than with small samples. Hence, they integrate findings across a number of primary research studies examining a conceptually similar hypothesis, by bringing them to a common metric called an effect size (Lipsey and Wilson, 2001).

In education, for example, a large number of primary research studies examines the effects of specific interventions on student achievement through experimental or quasi-experimental designs. Meta-analyses are particularly suitable for synthesizing effect sizes across a range of contexts reported in different studies examining a similar intervention, to estimate the cumulative effect size, the associated confidence intervals and the potential moderators of the effect size (Hattie *et al*., 2014).

Searching for relevant literature

In a meta-analysis, the criteria for inclusion and exclusion of primary research are explicitly stated. Comprehensive searches are conducted for relevant studies with explicit delineation of search protocols. While some meta-analysts argue that studies with relatively weak study designs ought to be excluded (Slavin, 2008), others insist that all research reports which meet the substantive selection criteria ought to be included in the synthesis, and an empirical examination should then be conducted of how different study design features moderate the effect size (Glass *et al*., 1981).

Meta-analysts have identified a number of publication biases and search biases along with strategies for taking these biases into account. A publication bias exists when the chances of a study being published depends on the nature of its methodological orientation or findings. A search bias exists when certain types of studies are more likely to be retrieved through common search channels, such as key databases. Studies with large sample sizes are more likely to attract research funding, being submitted for publishing and getting published in reputable journals. Research that does not report marked differences between individual groups or sub-groups examined within a study is less likely to be published (Rothstein *et al*., 2004). 'Sub-group reporting bias' exists when several sub-groups are compared but only comparisons with interesting or statistically significant findings get published. Similarly, 'time-lag bias' exists when certain types of studies, such as those with large sample sizes or effect sizes, take less time to get published (Sutton, 2009, p. 448). Unaccounted publication biases and search biases can lead to Type I error, leading to erroneous reporting of a large overall effect.

Extracting relevant information from selected studies

Meta-analysts follow explicit procedures for extracting relevant information from each study by developing protocols for coding contextual and outcome variables. Findings of individual studies are then converted to a common metric called an effect size, typically expressed as the difference of means between the experimental and control groups divided by the standard deviation. Algebraically,

$$(Me - Mc)/SD$$

where *Me* is the mean of the experimental group, *Mc* is the mean of the control group and *SD* is the pooled standard deviation. An effect size of $d=0.0$ is indicative of no change, while an effect size of $d=1.0$ is typically regarded as a blatantly obvious change (Cohen, 1988).

Meta-analytic literature contains sophisticated discussions of different types of effect sizes suitable for different study designs; formulae that allow conversion between different effect size indices; guidelines for estimating an effect size when some information is not reported in the primary research report; and appropriateness of various effect indices for analysing different types of data (Borenstein, 2009; Fleiss and Berlin, 2009).

Integrating findings across studies

Reasoning that findings from studies with large samples are more precise, often meta-analysts compute the composite effect by averaging the relevant d-statistics weighted by the reciprocal of their respective variances. Confidence intervals associated with the cumulative effect are also computed (Hedges and Olkin, 1985).

Within each category of conceptually similar effect sizes, the homogeneity statistic between the studies (Q_B) is estimated by assuming that Q_B has an approximate chi-square distribution with $m-1$ degrees of freedom, where m is the number of studies within each category. In rare cases, when the Q_B value is non-significant (indicating a consistency of outcomes across studies), the composite effect size is taken as a conclusive result representative of the within-category findings. However, often the Q_B value is significant, which indicates a considerable inconsistency across findings. In these cases, the composite effect size does not adequately describe the studies, since the magnitudes and perhaps the directions of the findings are very different from each other. These categories are analysed further to account for the differences in individual outcomes.

At this stage of the analysis, an outlier diagnosis is performed by visually plotting all the conceptually similar effect sizes to identify any effect size that markedly differs from the remaining effect sizes. If the study design of the outlier markedly differs from all the remaining studies, then the outlier is isolated and the difference in study design noted as a potential moderator of the effect size.

The remaining studies are subjected to categorical model testing, which is analogous to analysis of variance (ANOVA), to account for the heterogeneity of outcomes of different studies by identifying potential moderators of the effect. The studies are divided into sub-groups based on a study characteristic. Within each class, composite effect size and the within-group homogeneity statistic, Q_W, is estimated by assuming Q_W to have an approximate chi-square distribution with $k-1$ degrees of freedom, where k is the number of studies within each sub-group. A non-significant Q_W value indicates consistency of outcomes within a class. The

between-group homogeneity statistic (Q_B) is also estimated where a significant Q_B indicates that the study characteristic under consideration significantly moderates the effect size (Hedges and Olkin, 1985).

Meta-analysts sometimes conduct sensitivity analyses to 'assess robustness and bound uncertainty' (Orwin and Vevea, 2009, p. 196). In this procedure, the meta-analyst constructs connected understandings from the same data by making different assumptions about the data, including the missing data. The relative match between these constructed understandings demonstrates the degree to which these understandings are dependent on the initial assumptions made about the data. For instance, meta-analysts often compute inter-rater reliability in terms of multiple measures of inter-rater agreement; isolate outlier cases; isolate variables with low confidence ratings; use different formulae for computing or transforming effect sizes (e.g. Borman *et al*., 2003); and compare generalizations to the study sample based on a fixed effects model with generalizations to a large population based on a random effects model (e.g. Sirin, 2005). Meta-analysts employ various strategies to examine a potential publication bias (Sutton, 2009), such as: funnel plot method of plotting sample sizes of individual studies against their effect sizes; computing a fail-safe *n* (Rosenthal, 1980) to estimate the number of studies with insignificant findings that would have to be added to the analysis to make the cumulative effect size insignificant. Sensitivity analyses facilitate ratification and validation of conclusions geared for consensus and convergence.

There are several useful and cost-effective software programs for meta-analysis which offer many features including user-friendly menus to input information about an individual study's characteristics and data for computing effect sizes; compute and transform individual effect sizes; calculate cumulative effect sizes after appropriately adjusting relevant effect sizes; conduct homogeneity analyses, outlier diagnoses and categorical testing for identifying moderator variables; and conduct sensitivity analyses for comparing results based on different assumptions and analytic paths. See Shadish (n.d.) for a list of commonly used software for meta-analysis.

Presenting the findings as a scientific report

Meta-analysts typically employ the scientific reporting format with an explicit discussion of the critical decision points in the process. It typically has four distinct sections, i.e. Introduction, Methods, Results and Discussion. The meta-analyst begins with an identification and contextualization of the problem; describes their attempts to find the solutions; then they describe their findings using appropriate statistical and visual techniques; and finally they interpret and contextualize their findings (Cooper and Hedges, 2009).

21.3 Systematic reviews

In an age of evidence-based education, systematic reviews are increasingly used methods of investigation, bringing together different studies to provide evidence to inform policy making and planning (Gough *et al*., 2012). Several large centres have been established to support production and dissemination, for example:

- the EPPI-Centre (Evidence for Policy and Practice Information and Co-ordinating Centre) at the University of London (http://eppi.ioe.ac.uk/cms);
- the Social, Psychological, Educational and Criminological Controlled Trials Register (SPECTR) (Milwain, 1998; Milwain *et al*., 1999), later transferred to the Campbell Collaboration (www.campbellcollaboration.org), a parallel to the Cochrane Collaboration in medicine (www.cochrane.org), which undertakes systematic reviews and meta-analyses of, typically, experimental evidence in medicine;
- the Curriculum, Evaluation and Management (CEM) centre at the University of Durham (www.cemcentre.org);
- the What Works Clearinghouse in the US (http://ies.ed.gov/ncee/wwc);
- Centre for Reviews and Dissemination (www.york.ac.uk/crd); and
- Best Evidence Encyclopedia (www.bestevidence.org).

Like meta-analyses, systematic reviews use several techniques to minimize bias. They follow explicit protocols and criteria for searching for relevant primary, usually empirical studies, for example: their inclusion and exclusion; the standards for acceptable methodological rigour; their relevance to the topic in question; the scope of the studies included; team approaches to reviewing in order to reduce bias; the adoption of a consistent and clearly stated approach to combining information from across different studies; drawing careful, relevant conclusions and recommendations (Evans and Benefield, 2001).

In addition, systematic reviewers make noteworthy contributions to the methodology of research by embedding the following features in systematic reviews (Gough *et al*., 2012):

- according a greater agency of control to stakeholders in making decisions about how the synthesis should proceed;

- intention to regularly update their reviews with the new relevant studies in the field;
- intention to reduce duplication through explicit international collaborations;
- providing methodological support to groups interested in conducting systematic reviews;
- development of useful databases of intervention studies and systematic reviews to facilitate their dissemination and access; and
- utilizing technology strategically to update and disseminate relevant information widely.

There are two methodologically distinct perspectives prevalent among systematic reviewers. The first group is dominated by meta-analysts who recommend that ideally a systematic review must hold randomized controlled trials (RCTs) as the gold standard for individual studies (e.g. Campbell Collaboration, n.d.). The second group of systematic reviewers have engaged with various issues associated with including qualitative and mixed methods research in systematic reviews: developing efficient search strategies; appraisal criteria and synthesis techniques for research from different methodological traditions; engaging various stakeholders in formulating and critiquing their reviews' questions, protocols and summary reports; and stressing that systematic reviews must be complemented with other forms of reviews to facilitate informed decision making by different stakeholders (e.g. Petticrew and Roberts, 2006; Pope *et al*., 2007; Gough *et al*., 2012).

Since the mid-1990s, there has been a growing interest in systematic reviews of qualitative studies, especially in the areas of health care and public policy. A variety of methods have been proposed for synthesizing qualitative research from interpretive and critical-realist perspectives which vary along several dimensions (Barnett-Page and Thomas, 2009). Some methods have been developed to facilitate a fuller understanding of a phenomenon (Jensen and Allen, 1996) but others are aimed at generating mid-range theory (Eastabrooks *et al*., 1994; Zimmer, 2006) or 'lines-of-action' (Hannes and Lockwood, 2011a, p. 1633) to inform practical decision making. While some systematic reviewers recommend purposeful sampling for selecting studies, others recommend comprehensive searches and inclusion criteria. Some recommend including epistemologically similar qualitative studies in a synthesis; others recommend including studies from diverse epistemologies. And while many qualitative systematic reviewers recommend a grounded theory-like approach of axial coding for identifying themes emerging across studies, others note that a grounded theory-like approach to synthesizing research can sometimes become very resource-intensive and may not be viable. To improve efficiency, some systematic reviewers recommend starting with an a priori conceptual framework and modifying it as new themes emerge from the data (Carroll *et al*., 2011; Dixon-Woods, 2011).

Some systematic reviewers argue that in preference to quantitative or qualitative research syntheses, often 'mixed methods research syntheses' are more suitable for providing 'more complete, concrete and nuanced answers' to complex synthesis questions (Heyvaert, 2013, p. 671). Particularly noteworthy is Pawson's (2006) method of 'realist synthesis' to develop theory from successful as well as non-successful implementations of a programme. Rather than making global generalizations, realist reviews seek to explain how different aspects of a programme are likely to work in different circumstances. The reviewer begins by identifying the key theories underlying the specific phenomenon to formulate a more refined theory. Then the reviewer applies this theory successively to explain a number of successful and unsuccessful cases. With each application, the reviewer refines the theory. The salient features of Pawson's method include: purposeful sampling; including studies with diverse qualitative and qualitative designs; involvement of stakeholders in identifying the purpose; and tentative findings which inform decision makers of the likely implications of different decisions made in different situations rather than what works (Pawson, 2006).

21.4 Methodologically inclusive research syntheses

Meta-analyses and systematic reviews have made a substantial contribution towards advancing methods of research synthesis and have their own domains of applicability. Nonetheless, many systematic reviewers exclude a large proportion of research on the grounds of poor methodological quality, using evaluation criteria that are biased against certain paradigmatic orientations. Such an unacknowledged bias raises serious questions about the validity and generalizability of review findings (Pawson, 2006). Even in their inclusion of qualitative research, systematic reviewers often include only interpretive qualitative research and seek ideologically neutral evidence. The rhetorical effect of terms like 'evidence-based practice', 'systematic reviews', 'clarity', 'comprehensive', 'reliable', 'objectivity', 'replicable' not only might discredit opposition, but also may have the political impact of favouring post-positivism. Ironically, these key terms that are associated with systematic reviews are operationalized differently by different groups of systematic reviewers.

The problem here is not the subjectivity associated with these terms, but the systematic reviewers' denial of subjectivity itself (Hammersley, 2004).

Many systematic reviewers uncritically value objectivity and transparency of process, a priori protocols and exhaustive searches. Accordingly, advantages of emergent synthesis designs and purposeful sampling are less discussed in this body of literature. In reality, transparency itself is always subjective, partial and purposefully informed, where each way of showing is mirrored by a way of concealing, which may or may not be deliberate. Prescribing a priori rules to enhance objectivity, transparency and clarity could reduce the quality of reviews by discouraging reflection on important process decisions (MacLure, 2005; Kennedy, 2007). Hammersley (2003) observes that the phrase systematic reviews makes the other forms of reviews appear 'unsystematic', which can be misleading. Rather than categorizing reviews as 'systematic/unsystematic', he urges the educational research community to develop an appropriate typology for distinguishing between reviews with different foci (p. 5).

Recognizing the need to move beyond post-positivist reviews that often focus on constructing a coherent understanding of a field, there have been calls for reviews that challenge prevalent understandings, illuminate variations across contexts and highlight the tensions inherent in our understanding about a phenomenon (Eisenhart, 1998); reviews which reflexively interrogate the inequalities associated with educational research and practice (Meacham, 1998); and post-structural reviews that focus on constructing multiple understandings with a critical awareness that any understanding is inherently partial and situated within a particular perspective and time (Lather, 1999).

In critiquing the hegemonic dominance of systematic reviews, critical scholars alert us to several contentious issues, such as the problematics of formalization and systematization of research synthesis processes (e.g. Gallagher, 2004; MacLure, 2005), and power issues influencing how evidence is used (Clegg, 2005, p. 425). Many of these criticisms apply to most formal research synthesis methods and not just systematic reviews.

Suri (2014) draws upon the diverse literature on primary research methods and research synthesis methods to expand possibilities within research synthesis methods from a perspective of methodological inclusivity. Noting that contemporary educational research is marked by diversity, complexity and richness of purposes, methods and perspectives, she illuminates the variety of ways in which we can accommodate and reflect such variety and complexity at the level of synthesizing educational research.

Suri (2014) has identified the following three general guiding principles for a quality research synthesis:

1 informed subjectivity and reflexivity;
2 purposefully informed selective inclusivity;
3 audience-appropriate transparency.

Noting that each guiding principle will be enacted differently depending on the overarching epistemological and teleological orientation of the synthesis, Suri (2014) has clustered key decisions associated with the process of a research synthesis into six phases:

1 identifying an appropriate epistemological orientation;
2 identifying an appropriate purpose;
3 searching for relevant literature;
4 evaluating, interpreting and distilling evidence from selected reports;
5 constructing connected understandings;
6 communicating with an audience.

In the remainder of this chapter, key decisions associated with each of these phases of research synthesis are discussed from a methodologically inclusive perspective.

Phase one: identifying an appropriate epistemological orientation

There is no best-fit orientation for all research syntheses. The overarching orientation of the synthesis ought to be guided by the anticipated utility of the synthesis, the nature of primary research in the field and the synthesist's methodological expertise. The synthesist must attempt to make explicit the reflexive relationship between the synthesis findings and the synthesist's own research disposition. In Table 21.1, an illustrative framework of four epistemological orientations is used to demonstrate how syntheses with different paradigmatic orientations can serve varied, albeit equally useful, purposes.

The first row in Table 21.1 illustrates distinct ontological positions of a synthesis. The second row illustrates different purposes that a synthesis can serve. The third row illustrates potential relationships that a synthesist can have with participants and authors of primary research. Evidence included in a research synthesis is interpreted and represented first by the participants in primary research, then by the authors of primary research and finally by the research synthesist. Both participants and authors of primary research serve as informants for a research synthesist. The fourth row of Table 21.1 illustrates a range of strategies for searching and distilling relevant evidence and constructing

TABLE 21.1 RESEARCH SYNTHESES WITH DIFFERENT EPISTEMOLOGICAL ORIENTATIONS

	Post-positivist syntheses	*Interpretive syntheses*	*Participatory syntheses*	*Critical syntheses*
Ontological position	Objective factual world is out there	World is constructed through meanings that individuals and groups attribute to events	Individuals and groups construct their own world views through participation	Relativistic and transitional world views reflective of dominant power structures
Amenable purposes	Objectively explain, predict or describe in terms of probabilistic, generalizable laws, facts or relations between measurable constructs and variables	Construct deeper and more comprehensive understanding about phenomena as experienced subjectively by different stakeholders	Understand and/or improve ourselves and our local world experientially through critical engagement	Problematize prevalent metanarratives to deconstruct and/or transform dominant discourses
Informant–synthesist relationship	Objective distancing of an unbiased expert	Sensitive and reflective understanding with minimal power imbalance	Critical, selective and creative understanding, emphasizing realistic transferability to inform local practice	Self-doubting and reflexive understandings of perspectives represented in, and missing from, primary research literature
Common strategies	Exhaustive sampling; a priori protocol and coding sheets; statistical variable-oriented analysis	Purposive sampling; emergent design; holistic case-oriented analysis; summary sheets, meta-matrices, reciprocal translations, etc.	Purposive sampling; emergent design; eclectic data analysis; emphasis on practical and experiential knowledge	Openly ideological, dialogic, dialectic selection and analysis of evidence, emphasis on historical and structural insights
Quality criteria	Validity and reliability	Deep and authentic understanding	Empower participants to improve locally	Catalytic validity or crystallization
Suitable genres	Scientific reporting format	Comprehensive narrative with thick descriptions	Interactive reporting	Nuanced texts celebrating intertextuality

Source: Adapted from Suri (2013, p. 892)

connected understandings from the distilled evidence. The fifth row illustrates various quality criteria suitable for evaluating syntheses. The last row illustrates common genres that synthesists could employ to communicate with their audiences.

The four columns of Table 21.1 illustrate distinctions between research syntheses aligned with four distinct epistemological orientations: post-positivist syntheses; interpretive syntheses; participatory syntheses; critical syntheses. Boundaries between different orientations are blurred and rigid adherence to any single perspective is neither prescribed nor recommended. Nonetheless, synthesists should be critically aware of the implications of the choices they make, where some of these choices are likely to involve drawing from more than one paradigm.

Post-positivist syntheses: Often post-positivist synthesists, like meta-analysts, seek to synthesize research objectively with minimal researcher bias, by designing a priori synthesis protocols to minimize biases introduced by the synthesist's subjective preferences; defining conceptually and operationally all key constructs in behavioural terms at the outset; and employing exhaustive sampling in order to be representative of the entire population of studies. Sometimes post-positivist synthesists blind primary research reports to reduce biases introduced in judging the quality of individual reports by preconceived notions about the source of the publication or the author of the individual primary research report. Also, they measure inter-rater reliability to judge the degree of objectivity and reliability associated with the key decisions in the synthesis process

(Lipsey and Wilson, 2001; Orwin and Vevea, 2009; Wilson, 2009).

Post-positivist synthesists commonly employ variable-oriented statistical analyses to reduce Type II error and to enhance objectivity in the process of analysis and synthesis; target global decision makers and researchers as their audience; utilize the scientific reporting format; and adapt Cook and Campbell's (1979) constructs of validity and reliability to address issues of rigour in research synthesis. Sophisticated discussions have been published about ways of reducing threats to internal validity, external validity, internal reliability and external reliability within post-positivist syntheses (e.g. Petticrew and Roberts, 2006; Matt and Cook, 2009; see also Chapter 14 of the present volume).

Interpretive syntheses: In the last two decades, interpretive syntheses have been discussed under various names, such as meta-ethnography (Noblit and Hare, 1988), exploratory case-study oriented review of multivocal literatures (Ogawa and Malen, 1991), cross-case analysis (Miles and Huberman, 1994), aggregated analysis (Eastabrooks *et al*., 1994), meta-analysis of qualitative research (Jensen and Allen, 1994), qualitative meta-synthesis (Zimmer, 2006), interpretivist-oriented reviews (Eisenhart, 1998), meta-synthesis (Bair, 1999), meta-study (Paterson *et al*., 2001), thematic synthesis (Thomas and Harden, 2008) and framework synthesis (Carroll *et al*., 2011).

Contesting an objective reality that is out there, interpretive synthesists hold that the world is socially constructed in terms of the meanings we attribute to events. Typical questions addressed by an interpretive synthesist include: How do different stakeholders in different contexts experience a phenomenon? How do the contextual particularities interact with the perceptions of different groups and individuals? How do individual primary research reports on a topic reinforce, contradict or augment each other?

Interpretive synthesists begin by acknowledging the tacit knowledge, values and experiences that they bring to the synthesis process. They recognize that each primary research report is the author's interpretation of the research participants' interpretation of the phenomenon being studied. By engaging in iterative negotiations between multiple meanings constructed at each layer of interpretation and representation, they try to reveal the multiple perspectives of different stakeholders with a sensitive understanding. They seek evidence that contests, reinforces or augments their emerging understanding of the phenomenon. Refraining from the tendency to construct one coherent grand metanarrative, they remain open to constructing multiple understandings of the phenomenon (Paterson *et al*., 2001).

Participatory syntheses: Participatory synthesists encourage critical thinking through engaged participation with those whose practices and experiences are being researched. A complementary collaborative model, where the distinct skill sets and expertise of individual collaborators are valued, is suitable for encouraging participation of different stakeholders in a research synthesis, without burdening them with a heavier workload (Ritchie and Rigano, 2007; Yu, 2011). Rather than ironing out the differences, a participatory synthesis involves paying careful attention to learning opportunities that arise from the differences in language, perspectives and experiences of individual co-synthesists (Paugh and Robinson, 2011). A participatory synthesis can become a site for teachers where they collaboratively reflect on their own practice using published research as a mirror to develop 'actionable knowledge' about their own practice (Torrance, 2004, p. 198).

Participatory synthesists value practical experience, local knowledge and serendipitous leaps of intuitive understanding. Such participants could be: the authors of the primary research reports being synthesized; members of stakeholder groups who participated in those studies; or stakeholders wishing to engage critically with the literature to inform their own decisions. Academic synthesists collaborate with these participants in order to co-synthesize the relevant body of research through a process of reciprocal learning and co-constructing connected understandings. A participatory synthesis of action research reports authored by teacher-researchers or reflexive practitioners, on how they effected changes within their contextual constraints, can provide useful information to policy makers and other practitioners. Identifying patterns across these individual reports can provide useful input from this group of action researchers towards theory-building.

A participatory synthesis can involve: cycles of reflection to formulate synthesis purpose; conducting the research synthesis; implementing changes as suggested by the implications of the synthesis; evaluating the implemented change and comparing these evaluations with the relevant research literature. Using emergent, pragmatic and eclectic designs, participatory synthesists can employ purposeful sampling strategies for selecting studies which illuminate aspects of a phenomenon that are of immediate interest to the participant co-synthesists. They can use the delphi-technique for collecting, analysing and building collective understandings of research to involve a homogeneous or heterogeneous group of participants with a common interest in the research topic. Participatory synthesists can construct critical, selective and creative understandings with

realistic transferability to inform practice in local contexts of the participants. They can employ an interactive reporting format to encourage a participative audience. The synthesis can be evaluated in terms of the progress in thinking and transformation of the contexts of the individuals and the communities of those engaged in the synthesis process (see also Chapter 3 of the present volume). A practical example of some of these ideas can be seen in the work of Bassett and McGibbon (2012), who designed a critical, participatory and collaborative method for scoping literature with individuals who had been actively drawing attention to barriers to health and well-being in rural, Aboriginal and African Canadian communities in Canada.

Critical syntheses: Critical synthesists can reveal how certain forms of knowledge get privileged and regulate the norms within discussions of policy, practice and research in a field (Aronowitz and Giroux, 1991; Clegg, 2005; see also Chapter 3 of the present volume). By paying attention to the presence and absence of various issues in the primary research reports, critical synthesists can disrupt conventional thinking to construct spaces for new ways of talking about policy, practice and research (Eisenhart, 1998; Segall, 2001; Kress, 2011). A good example of a critical synthesis is Windschitl's theoretical analysis of research to illuminate the uncertainties and tensions experienced by teachers along with the compromises they make as they implement a constructivist pedagogy in practice (Windschitl, 2002, p. 131).

Through critical interrogation of the very text being synthesized and by constantly questioning the assumptions we make in constructing our understanding, critical synthesists could open new ways of thinking about a phenomenon. They can challenge the prevalent understanding by revealing the errors and elements of ignorance underpinning the prevalent understanding of a phenomenon. Rather than forcibly constructing a single coherent understanding of a phenomenon, critical synthesists could illuminate multiple understandings of a phenomenon to expand possibilities for practitioners (Sholle, 1992). Postmodern synthesists, as a sub-group of critical synthesists, could disrupt and problematize the metanarratives in a research domain in order to enhance multiple discourses that celebrate diversity and inclusivity by refusing to provide simplistic explanations for complex phenomena (Lather, 1993).

Examples of questions addressed by critical synthesists could include the following: What are the gaps in our understanding of a phenomenon? What methodologies or theoretical perspectives are likely and/or unlikely to be employed by primary researchers in the field? In the published literature, whose questions are prioritized? Whose questions have received little attention from primary researchers? How are the answers to such questions intertwined?

Critical synthesists could construct self-doubting and reflexive understandings of not only the perspectives represented in the primary research literature but also those missing from the published primary research to illuminate how some groups have become invisible in the field with little representation. Critical synthesists can also collaborate with the groups who have been relatively silenced in the primary research in order to identify how the body of primary research has failed to adequately represent their interests (see, e.g., Warschauer and Matuchniak, 2010). Rather than deferring to the *authority of the author*, postmodernist critical synthesists would recognize an author as someone who is in the process of making sense, a sense which is partial and temporal (Lather, 1999; Richardson, 2001).

Phase two: identifying an appropriate purpose

An emphasis on purposefully informed selective inclusivity necessitates that synthesists carefully identify a purpose that takes into account:

- potential stakeholders and collaborations;
- the nature of the substantive area;
- the intended audience and utility;
- pragmatic constraints;
- ethical considerations.

All these factors might influence, and be influenced by, the synthesist's contextual positioning and the overarching epistemological, theoretical and political orientations of the synthesis.

A number of groups can differentially influence or be affected by a research synthesis, including the following: learners, families of learners, educators and educational institutions, primary researchers in the substantive area, policy makers, the wider community, commercial and political groups with an interest in the topic. Funding agencies, editorial boards and communities, and professional synthesists can also feature in these groups.

Stakeholders in a synthesis are often anticipatory rather than retrospective. Synthesists can purposefully collaborate with diverse stakeholders to achieve some of the following objectives:

- encourage syntheses that address the concerns of a wide range of stakeholders;
- facilitate syntheses informed by the perspectives of different groups;

- empower members of different groups by facilitating their participation in syntheses which may be of interest to them;
- enhance the impact of a synthesis by promoting participation of the agents of change who are crucial in implementing the recommendations made by the synthesis;
- contribute to wider dissemination of research syntheses;
- deepen academic synthesists' understandings of the collaborating stakeholders' concerns and understandings.

Different collaborators have the potential to enrich the synthesis by bringing in their own particular expertise. Each form of collaboration also introduces issues of power and varied interests that can add complexity to the synthesis process (Yu, 2011). In a collaborative synthesis, synthesists must carefully negotiate issues arising from different perceptions of roles, responsibility, collaboration, authority and authorship (Baldwin and Austin, 1995).

When seeking input from stakeholders, synthesists should sensitively clarify the nature of input being sought and what can or cannot be negotiated; address power imbalances between different stakeholders; recognize heterogeneity within stakeholder groups; and ensure that less powerful groups do not feel further disempowered with the perception that their views are not being paid adequate attention (Petticrew and Roberts, 2006; Rees and Oliver, 2012).

Often, research synthesists begin by reading previous research reviews in the field. Previous reviews, along with their bibliographic references, can provide useful information for developing a broad overview of primary research and research syntheses reported in the field. In formulating an appropriate synthesis purpose, synthesists often consider: the topicality of the field; the nature of predominant methodologies employed in the primary research studies; the general relationship between individual studies; and the volume and scope of the relevant primary research. Paying careful attention to ethical issues of representation and non-representation, synthesists should formulate the purpose for an intended audience and synthesis goal.

Phase three: searching for relevant literature

Research synthesists draw their evidence from the primary research, secondary research and previous research syntheses reported in a field. Research syntheses on the same topic conducted for different purposes can have different sampling strategies, each being equally legitimate but tailored to serve different purposes. Synthesists must search strategically for the relevant evidence to meet the synthesis purpose efficiently within the available resources and pragmatic constraints.

Several publication biases and search biases can influence funding, publishing and visibility of certain types of primary research as well as research synthesis (Petticrew and Roberts, 2006). Funding bias, methodological bias, outcome bias and confirmatory bias are examples of publication bias. Database bias, citation bias, availability bias, language bias, country bias, familiarity bias and multiple publication bias are examples of search bias that must be considered in planning appropriate search techniques. Through a careful cost–benefit analysis, synthesists make decisions related to inclusion or exclusion of unpublished reports.

The terms 'exhaustive' and 'expansive' are sometimes used to distinguish between two approaches to searching for suitable studies in recent literature on qualitative evidence synthesis in health care (Finfgeld-Connett and Johnson, 2012). Exhaustive searches are more suitable for integrative syntheses aimed at producing generalizable findings. They are typically employed by meta-analysts and systematic reviewers. Expansive searches with purposeful sampling strategies are more suitable for syntheses with emergent designs, where the search criteria evolve as the synthesis progresses, and are aimed at facilitating understanding, participation, emancipation or deconstruction (Suri, 2011).

A clear set of inclusion criteria defines the scope of the synthesis. The level of specificity associated with the inclusion criteria at different stages of the synthesis may vary in accordance with the synthesis purpose. The synthesist must strategically choose and sequence the use of appropriate search channels in a way that is aligned with the sampling logic and which yields the most relevant, trustworthy and comprehensive evidence within the available resources.

Phase four: evaluating, interpreting and distilling evidence from selected studies

While some synthesists argue that all research reports which meet the substantive selection criteria ought to be included in the synthesis (e.g. Jensen and Allen, 1996; Glass, 2000), others insist that studies which have relatively weak study designs ought to be excluded (e.g. Eastabrooks *et al*., 1994; Slavin, 2008). All synthesists agree that findings from primary research that have stronger study designs should be given more weight in constructing collective understandings. Rather than asking 'Is this a perfect study?',

a synthesist ought to ask 'How do methodological features of this study impact upon the trustworthiness of its findings in ways that are relevant to my synthesis purpose?' Synthesists must examine each report closely for a coherence between its theoretical background, intended purpose, context and/or nature of intervention being studied, methods for collecting, analysing and interpreting evidence, results and conclusions. Equally important is the relevance of the report for the synthesis purpose (Dixon-Woods *et al*., 2006; Major and Savin-Baden, 2010).

In a rigorous synthesis, the set of evaluation criteria for individual studies is essentially guided by the overarching teleological and theoretical orientation of the synthesis. For example, studies with representative samples are suitable for generating generalizations. However, issues of how the views and voices of different stakeholders are represented become particularly important in an emancipatory synthesis. Useful discussions have been published for evaluating the quality of primary research reports stemming from different traditions, such as those for post-positivist research (e.g. Centre for Reviews and Dissemination, 2009; Valentine, 2009), interpretive research (Hannes *et al*., 2010), participatory research (Heron and Reason, 1997) and critically oriented research (Lather, 1986).

Ongoing reflexive engagement with the selected studies is crucial to reduce unaccounted or unacknowledged biases. While coding information from individual studies, synthesists can assign 'confidence ratings' to findings or insights that require higher levels of interpretation (Stock, 1994, p. 128). One approach to coding can be to assign each finding one of the following three codes based on the degree to which the finding is supported by reported evidence: compelling, credible and unsupported (Major and Savin-Baden, 2010). Decisions in relation to dealing with missing data or biased findings should be consistent, substantiated and disclosed.

To minimize unstated subjective biases, meta-analysts and systematic reviewers try to maintain consistency and transparency by adhering to a priori protocols. Research synthesists espousing critical orientations, on the other hand, stress that synthesists should reflexively respond to change, rather than rigidly follow an a priori protocol (Zhao, 1991; Pawson *et al*., 2005). A reflexive stance involves constantly reflecting critically on how the synthesist's own dispositions and perspectives are dialectically influencing, and being influenced by, the synthesis process and findings (MacLure, 2005).

Several authors recommend that the complexity of a research synthesis process requires collaborative efforts to ensure a certain level of trustworthiness: to minimize subjective judgements and biases by maximizing inter-coder reliability (Stock, 1994) or by deliberately critiquing and contesting each other's emerging understandings (Ogawa and Malen, 1991); as a form of triangulation to improve rigour (Eastabrooks *et al*., 1994); or for co-constructing collective interpretations through dialogic discussions (Wideen *et al*., 1998, p. 135; Paterson *et al*., 2001, p. 47).

Phase five: constructing connected understandings

While some research synthesis methods focus on variable-oriented connections, others focus on study-oriented connections. In variable-oriented connections, the variations in the effects, implementations, manifestations, meanings, understandings or conceptions of a phenomenon are the prime focus of a synthesis. Each account is examined to the extent that it contributes to explaining the relationships between the target variables. When several individual studies examine a particular intervention, concept or phenomenon using similar methodologies, the findings may be aggregated to increase size and variations of the overall sample. These findings can then be used to make generalizations, provide plausible explanations and predict patterns of human behaviour as in a meta-analysis or an aggregated analysis (Eastabrooks *et al*., 1994). Comparative techniques such as content analysis, statistical techniques and visual displays are commonly employed for constructing variable-oriented connections.

Study-oriented connections are amenable for constructing holistic and complex understandings with an attempt to retain the contextual integrity of individual accounts, as in Noblit and Hare's (1988) meta-ethnography. The focus here is on understanding the dynamics of individual accounts as the synthesist attempts to reconcile the dynamics of each study with those of the other studies. The synthesist tries to determine the relations and tensions between individual accounts through a dialectical process of comparing key constructs, phrases and themes used in each study as interpreted by each co-synthesist (Noblit and Hare, 1988).

When individual reports are addressing similar issues, they are amenable to a reciprocal translational synthesis to construct a collective understanding that captures the essence of all the included studies (Jensen and Allen, 1996; Paterson *et al*., 2001). The findings of each report are tested for their abilities to translate the findings of the other reports. Those terms or findings are selected which can more succinctly describe the findings of all the reports within the subset. When the synthesist has limited resources, the synthesist can

select an exemplary study and examine other studies for the extent to which they demonstrate or add to the description of the phenomenon in the exemplar study (Miles and Huberman, 1994).

When individual reports give conflicting representations of the same phenomenon, they lend themselves to a 'refutational synthesis' (Noblit and Hare, 1988, p. 47) where the relationships between individual studies and the refutations become the focus of the synthesis process. The contradictions between individual reports may be explicit or implicit. The implicit refutations are made explicit using an interpretive approach. This is followed by an attempt to explain the refutation.

If individual reports examine different aspects of the same phenomenon, a 'lines-of-argument' synthesis could be used to make inferences (Noblit and Hare, 1988). In this method, findings from individual reports are used as pixels to get a fuller picture of the phenomenon at hand. It involves a grounded theory-like approach for open-coding and identifying the categories emerging from the data. The key categories that are more powerful in representing the entire data set are identified by constant comparisons between individual accounts. These categories are then linked interpretively to create a holistic account of the phenomenon.

When synthesizing methodologically diverse reports, a synthesist can begin by constructing collective understandings from clusters of studies with similar designs and then synthesize collective understandings across clusters (Suri, 1999; Greenhalgh *et al*., 2005).

Study-oriented connections can also be implemented, as in Pawson's realist synthesis, by inferring theory from each study and examining its transferability to other cases to refine the initial theory. In a realist synthesis, this process is repeated with every study of successful and unsuccessful implementation to develop a more sophisticated and comprehensive theory that can explain many cases (Pawson *et al*., 2005).

Strategies for enhancing plausibility, authenticity, utility, robustness and validity of synthesis findings include:

- reflexivity;
- collaborative sense-making;
- eliciting feedback from key stakeholders;
- identifying disconfirming cases and exploring rival connections;
- sensitivity analyses and using multiple lenses to identify the dependence of a synthesis finding on underlying assumptions or the frame of reference.

Phase six: communicating with an audience

It is crucial that synthesists carefully select the content, representational style, medium, genre and techniques in alignment with the impact they wish to make on their target audience. They must share the process and the product of their sense-making skilfully and in a way that is credible, trustworthy and useful to the intended audience. A range of interesting, engaging and effective possibilities for communication can arise by embedding various representational tools within diverse structural genres expressed through varied media. As educational reviews are often seen to represent wider power relations, synthesists should be critically conscious of how a review may preferentially normalize certain representations or practices over the others (Baker, 1999).

Many synthesists, especially meta-analysts, employ the scientific reporting format with four distinct sections, i.e. Introduction, Methods, Results, Discussion. Many others use coherent thematic narratives to share their synthetic understandings with an audience. They construct their narrative by organizing into a coherent logical structure the main themes which contextualize and describe the synthesis process and product. These themes, headings and subheadings may vary purposefully according to: intervention types; methodological designs; contextual features (e.g. Engberg, 2004); types of descriptive commentaries such as retrospective, prospective or critical (e.g. Bransford and Schwartz, 1999); or the aspect of the synthesis process being described such as introduction, methods, results and discussion.

For each idea, theme or finding, synthesists should identify pertinent techniques for a rich, succinct and audience-friendly representation. Given the vast nature of evidence in research syntheses, descriptive statistics, abridged quotes and visual displays can be particularly useful tools. Synthesists can also choose from a range of narrative techniques and artistic devices.

Quality research synthesis reports generally share the following characteristics: conceptually substantiated and well-bound coverage of the substantive topic; rigorous critique of previous reviews; identification of common assumptions, theories, methods and findings emerging from extant research; critical analyses of extant research; coherent structuring of the report along meaningful themes; a unique conceptual framework or perspective to think about the topic, future research, practice and policy; clear implications for researchers, practitioners and policy makers; and a discussion of any caveats associated with synthesis findings to clarify its domain of applicability (Suri, 2014).

21.5 Conclusion

This chapter has discussed many of the key issues in carrying out meta-analysis, systematic review and methodologically inclusive research syntheses as an educational researcher. The chapter's central focus settles on the many critical decisions associated with each of six phases of a research synthesis, and these are discussed from a methodologically inclusive perspective. The strengths, complexities, domains of applicability and caveats of the different approaches have been discussed so that researchers can make their own informed choices about the kind of research approach they wish to adopt for a given project.

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Action research

CHAPTER 22

Action research is a widely used approach to research, as evidenced in the numerous journals whose titles bear the phrase 'action research'. The main aim of action research, as one of its key proponents, Elliott (1991), states, is 'to improve practice' (p. 49). It is a methodology for researchers (often teachers) to understand and generate knowledge about educational practices and their complexity (McAteer, 2013, p. 21).

In action research, typically teachers and other parties research their own institutions, staff development facilitators bring about change and groups and communities undertake research. This chapter introduces key issues in the planning, conduct and reporting of action research, including:

- defining action research
- principles and characteristics of action research
- participatory action research
- action research as critical praxis
- action research and complexity theory
- procedures for action research
- reporting action research
- reflexivity in action research
- ethical issues in action research
- some practical and theoretical matters

The chapter draws links between action research and critical theory, in particular in respect of participatory action research. It also notes the connections between action research and complexity theory.

22.1 Introduction

Action research is a useful tool for change and improvement at the local level. Indeed Kurt Lewin's own work (one of action research's founding fathers) was deliberately intended to change the life chances of disadvantaged groups in terms of housing, employment, prejudice, socialization and training. Its combination of *action* and *research* has contributed to its attraction to researchers, teachers and the academic and educational community, breaking the culture of 'spectator research' (Cain, 2011, p. 3) and 'ivory tower' research (Munn-Giddings, 2012, p. 71). Action research, as one of its leading proponents, Stephen Kemmis (2009) notes, seeks to change and transform 'practitioners' *practices*, their *understandings* of their practices, and the *conditions* in which they practice' (p. 463). It is a '*practice-changing practice*' (p. 464; italics in original; see also Kemmis *et al*., 2014). McAteer (2013) adds to this the point that, by being a collaborative process, it aims to become both a 'democratic and democratising process' (p. 17). Cain (2011) argues that action research moves beyond positivist, interpretive and critical research, being self-reflexive, collaborative, political and suitable for dissemination, for example, to other teachers (pp. 13–14). McNiff (2010) advises researchers to

> take action to improve something; make sure you understand what you have done; use understanding (knowledge) to give explanations for how and why you have improved it (in academic terms, this means that you generate new theory – the word 'theory' means explanation).
>
> (McNiff, 2010, p. 16)

Action research, she observes, is 'practice-based' (p. 33), concerns learning and the creation of knowledge, is 'values laden', 'educational', 'collaborative', 'critical and risky' and 'always political' (pp. 33–4), and she provides plentiful advice for researchers working in this mode.

Action research can be used in almost any setting where a problem involving people, tasks and procedures needs a solution, or where some change of feature results in a more desirable outcome. It embraces both problem posing and problem solving. It need not focus solely on problems; it can also embrace areas of interest for development (McNiff, 2010). It can be undertaken by the individual teacher, a group of teachers working cooperatively within one school or a teacher or teachers working alongside a researcher or researchers in a sustained relationship, possibly with other interested parties like advisers and university departments (Holly and Whitehead, 1986). Ferrance (2000)

identifies different levels of action research: individual teacher research, collaborative action research, school-wide action research and district-wide actions research (p. 6).

Action research can be used in a variety of areas, for example:

- *teaching methods*: replacing a traditional method by a discovery method;
- *learning strategies*: adopting an integrated approach to learning in preference to a single-subject style of teaching and learning;
- *evaluative procedures*: improving one's methods of continuous assessment;
- *attitudes and values*: encouraging more positive attitudes to work, or modifying pupils' value systems with regard to some aspect of life;
- *continuing professional development of teachers*: improving teaching skills, developing new methods of learning, increasing powers of analysis of heightening self-awareness;
- *management and control*: the gradual introduction of different techniques of class management;
- *administration*: increasing the efficiency of some aspect of the administrative side of school life.

It can be used to foster democratic institutions, encourage change, empower individuals and groups, encourage reflective practice and be a test-bed for new ideas and practices (Creswell, 2012, p. 578).

These examples do not mean, however, that action research can be typified straightforwardly; that is to distort its complex and multifaceted nature. Indeed Kemmis (2010) suggests that there are several schools of action research. That said, what unites different conceptions of action research is the desire for improvement to practice, based on a rigorous evidential trail of data and research.

Ferrance (2000, p. 1), echoing Watts (1985), argues that a powerful justification for action research is that teachers:

- work best on problems that they have identified for themselves;
- become more effective when they are encouraged to examine and assess their own work and then consider ways of working differently;
- help each other by working collaboratively;
- help each other in their professional development by working together.

She suggests that action research builds on, and builds in, these principles. Indeed action research is a potent form of participatory research (see Chapter 3; Kapoor and Jordan, 2009). 'Commitment' is a feature of participatory research (see Chapter 3), and participatory action research both requires and builds commitment (David, 2002). Participatory research breaks the separation of the researcher and the participants; power is equalized and, indeed, they may all be part of the same community. The research becomes a collective and shared enterprise in many spheres, including: research interests, agendas and problems; generation and analysis of data; equalization of power and control over the research outcomes, products and uses; development of participant voice, authorship and ownership; emancipatory agendas and political goals; a process-orientated and problem-solving approach; and ethical responsibility and behaviour. We refer the reader to Chapter 3 for a fuller discussion of this.

22.2 Defining action research

Action research comes in many forms, including action research, participatory action research, critical action research, diagnostic action research, practitioner research, classroom-based action research, empirical action research and many others (e.g. Jefferson, 2014). It is typically a small-scale intervention in the functioning of the 'real' world and a systematic, close examination, monitoring and review of the effects of such an intervention, combining action and reflection to improve practice (cf. Ebbutt, 1985, p. 156; Hopkins, 1985, p. 32; New South Wales Department of Education and Training, 2010). When operating with collaborative groups of teachers, it is designed to address collective improvement and development. Action research is an investigation which is intentionally directed towards solving a problem or focusing on an issue raised by, and owned by, an individual or a group (Kemmis and McTaggart, 1988, p. 5; McNiff, 2010). It has a deliberately applied focus (Creswell, 2012, p. 577) and is 'hands on' research (Denscombe, 2014, p. 122).

Piggot-Irvine *et al*. (2015) define action research as 'a collaborative transformative approach with joint focus on rigorous data collection, knowledge generation, reflection and distinctive action/change elements that pursue practical solutions' (p. 548). This draws attention to two key points taken up further in this chapter: the collaborative, emancipatory claims of action research, and its practical outcomes in terms of bringing about change.

The rigour of action research is attested by one of its founding fathers, Corey (1953, p. 6), who argues that it is a process in which practitioners study areas for development *scientifically* (our italics) so that they can

evaluate, improve and steer decision making and practice. (Helskog (2014) discusses the debate as to whether or not action research is a science and what the implications of this debate are in regard to justifying action research.) Kemmis and McTaggart (1992, p. 10) argue that action research requires systematic planning, acting, observing and reflecting in a manner that is more demanding and rigorous than in the everyday course of life.

A philosophical stance on action research that echoes the work of Habermas is taken by Carr and Kemmis (1986, p. 162), who regard it as a form of 'self-reflective enquiry' by participants which is undertaken in order to improve their understanding of their practices with a view to maximizing social justice (see also Gibbs *et al*., 2016). Grundy (1987, p. 142) regards action research as concerned with improving the 'social conditions of existence'. Kemmis and McTaggart (1992) suggest that action research is concerned with changing both individuals and the institutions, societies and cultures of which they are members.

Action research is designed to bridge the gap between research and practice (Somekh, 1995, p. 340), thereby striving to overcome the perceived persistent failure of research to impact on, or improve, practice. Stenhouse (1979) suggests that action research should contribute not only to practice but to a theory of education and teaching which is accessible to other teachers, making educational practice more reflective (Elliott, 1991, p. 54).

Action research combines diagnosis, action and reflection (McNiff, 2010), focusing on practical issues that have been identified by participants and which may somehow be both problematic yet capable of being changed (Elliott, 1978, pp. 355–6). McNiff (2010) places self-reflection at the heart of action research, suggesting that whereas in some forms of research, the researcher 'does research on other people', in action research, the researcher does it to herself/himself. Jefferson (2014) notes that key assumptions of action research include that 'practitioners work best on problems that they have identified themselves' and that they increase their effectiveness if they examine and 'assess their own work and then consider ways of working differently' (pp. 91–2) and work collaboratively, thereby developing their professional practices. Ferrance (2000) and McNiff (2010) note that it is not concerned simply with solving problems ('what is wrong'; p. 2) but is more concerned with how to improve. As Zuber-Skerritt (1996b, p. 83) remarks: 'the aims of any action research project or program are to bring about practical improvement, innovation, change or development of social practice, and the practitioners' better understanding of their practices'.

The several strands of action research are drawn together by Kemmis and McTaggart (1988) in their view of action research as a type of critical reflective enquiry which participants undertake on and for themselves, focusing on problems and practices which they identify themselves and which affect them, with the intention of understanding and improving the educational and social practices in which they are involved and the circumstances in which they take place, in order to promote social justice; it is undertaken collaboratively, albeit sometimes focusing on individuals in the group (p. 5).

Kemmis and McTaggart (1992, pp. 21–2) distinguish action research from the everyday actions of teachers: in four main ways:

- it is thinking in a more systematic and collaborative way than the customary, everyday ways in which teachers consider their own practices;
- it moves beyond problem solving alone to identifying and raising problems, regarding problem solving as opportunities for change, learning and improvement rather than as simply curing ills (a pathological model);
- it involves participants working on their self-identified areas of work, i.e. it is owned by the participants rather than external researchers;
- it adopts a heterogeneous rather than unitary concept of the science of teaching.

Noffke and Zeichner (1987) make several claims for action research with teachers, namely that it: brings about changes in their definitions of their professional skills and roles; increases their feelings of self-worth and confidence; increases their awareness of classroom issues; improves their dispositions towards reflection; changes their values and beliefs; improves the congruence between practical theories and practices; and broadens their views on teaching, schooling and society.

Action research lays claim to the professional development of teachers; action research for professional development is a frequently heard maxim (e.g. Somekh, 1995, p. 343, 2006; Winter, 1996; Ferrance, 2000; New South Wales Department of Education and Training, 2010). It is 'situated learning': learning *in* the workplace, *about* the workplace and *for* the workplace (cf. Collins and Duguid, 1989; NSW Department of Education and Training, 2010). The claims for action research, then, are several. Arising from these claims and definitions are several principles.

22.3 Principles and characteristics of action research

Hult and Lennung (1980, pp. 241–50), McKernan (1991, pp. 32–3), Ferrance (2000), the New South Wales Department of Education and Training (2010) and Kemmis *et al*. (2014) suggest that action research:

- makes for practical problem posing and problem solving as well as expanding scientific knowledge;
- enhances the competencies of participants;
- is collaborative;
- is undertaken directly *in situ*;
- uses feedback from data in an ongoing cyclical process;
- seeks to understand particular complex social situations;
- seeks to understand the processes of change within social systems;
- is undertaken within an agreed framework of ethics;
- seeks to improve the quality of human actions;
- focuses on those problems that are of immediate concern to practitioners;
- is participatory;
- frequently uses case study;
- tends to avoid the paradigm of research that isolates and controls variables;
- is formative, such that the definition of the problem, the aims and methodology may alter during the process of action research;
- includes evaluation and reflection;
- is methodologically eclectic;
- contributes to a science of education;
- strives to render the research useable and shareable by participants;
- is dialogical and celebrates discourse;
- has a critical purpose in some forms;
- strives to be emancipatory.

Zuber-Skerritt (1996b, p. 85) suggests that action research is: *critical* (and self-critical) collaborative inquiry by *reflective* practitioners being *accountable* and making results of their enquiry public, *self-evaluating* their practice and engaging in *participatory* problem solving and continuing professional development. This is echoed in Winter's (1996, pp. 13–14) six key principles of action research:

1. *reflexive critique*, which is the process of becoming aware of our own perceptual biases;
2. *dialectical critique*, which is a way of understanding the relationships between the elements that make up various phenomena in our context;
3. *collaboration*, which is intended to mean that everyone's view is taken as a contribution to understanding the situation;
4. *risking disturbance*, which is an understanding of our own taken-for-granted processes and willingness to submit them to critique;
5. *creating plural structures*, which involves developing various accounts and critiques, rather than a single authoritative interpretation;
6. *theory and practice internalized*, which is seeing theory and practice as two inter-dependent yet complementary phases of the change process.

Action research has been compared to, and claimed to differ from, 'formal research' in several respects (New South Wales Department of Education and Training, 2010, p. 1), for example: it requires little training (whereas formal research often requires extensive training); its intention is to focus on a local, institutional situation or practice with the intention of improving practice, in contrast to those kinds of research which seek generalizability; the sample comprises the relevant participants in the institution or situation rather than representing the wider population or being drawn randomly from the wider population; and it is of practical rather than theoretical significance.

Action research resists the criticism that it is not 'proper research' (Somekh, 2006), as it abides by many of the tenets of research (e.g. hypothesis generation and testing, new ways of looking at problems, generating knowledge, rigorous investigation) and, indeed, it benefits from insider knowledge. Similarly, Casey (2013) notes that action research benefits from being conducted by 'someone native to the field' and with action taken 'at the site where practice occurs' (p. 149). Teachers, for example, do it to and for themselves, becoming part of their everyday practices (p. 149).

The several features that the definitions at the start of this chapter have in common suggest that action research has key principles. These are summarized by Kemmis and McTaggart (1992, pp. 22–5), when they aver that action research:

- seeks to improve education by deliberately and deliberatively trying to change it, focusing on improving participants' own practices collaboratively and with consequent learning from these changes and their involvement in the process;
- involves ongoing, systematic and ongoing developmental cycles of planning, implementing, observing and reflecting (self-reflection) on changes. Such actions contribute to collaborative communities of

practitioners which are characterized by their self-critical reflection on, and insights into, relationships between contexts, actions and outcomes, and which promote emancipation, empowerment, legitimation and social justice in participants', others' and institutions' educational values and practices, thereby constituting a political process and endeavour;
- requires participants to enquire into and theorize about the conditions of, and circumstances and practices in, their own lives. In this process, their ideas, values and assumptions are tested against rigorous evidence in an open-minded, evidence-based, critical-analytical and reflexive spirit;
- leads to the development of a clearly justified, justifiable, reasoned, evidence-based rationale for educational practices;
- can begin small-scale, even on an individual or small-group level, and can spiral out to affect and involve others in communities of practice, thereby engaging issues of power and empowerment in decision making.
- can benefit from careful records of changes and improvement, the processes involved in making them, ways of perceiving and understanding them, the social relationships involved in them, and raised awareness of constraints on situations and participants.

Though these principles find widespread support in the literature on action research, they require some comment. For example, there is a strong emphasis in these principles on action research as a cooperative, collaborative activity (rather than an individualistic activity) (Kemmis and McTaggart, 1992, p. 15). Indeed Kemmis and McTaggart (1992) locate this in the work of Lewin himself, commenting on his commitment to group decision making (p. 6). They argue, too, that it is those groups who are involved in, or affected by, planned interventions and changes who should bear the prime responsibility for taking decisions on actions, based on reasoned, interrogated, evidence-based analysis and evaluation (p. 15).

The view of action research solely as a group activity, however, might be too restricting. It is possible for action research to be an individualistic matter as well, relating action research to the 'teacher-as-researcher' movement (Stenhouse, 1975; Pring, 2015). Whitehead (1985, p. 98) explicitly writes about action research in individualistic terms, and we can take this to suggest that a teacher can ask herself or himself: 'What do I see as my problem?' 'What do I see as a possible solution?' 'How can I direct the solution?' 'How can I evaluate the outcomes and take subsequent action?'

The adherence to action research as a group activity derives from several sources. *Pragmatically*, Oja and Smulyan (1989, p. 14), in arguing for collaborative action research, suggest that teachers are more likely to change their behaviours and attitudes if they have been involved in the research that demonstrates not only the need for such change but that it can be done – the issue of 'ownership' and 'involvement' that finds its parallel in management literature which suggests that those closest to the problem are in the best position to identify it and work towards its solution (e.g. Morrison, 1998).

Ideologically, there is a view that those experiencing the issue should be involved in decision making, itself hardly surprising given Lewin's own work with disadvantaged and marginalized groups, i.e. those groups with little voice (cf. David, 2002).

Politics and ideology are brought together in *participatory action research* and *action research as critical praxis*, and it is to this that we turn.

22.4 Participatory action research

Some researchers differentiate action research from participatory action research, the latter being a more specific subset of action research, whilst other researchers make no such distinction (Munn-Giddings, 2012, p. 72). Participatory action research has attracted attention across the world in its advocacy of democracy, empowerment and emancipation (cf. David, 2002; Jones and Stanley, 2010). Kemmis and McTaggart (2005) comment that participatory action research seeks to create conditions for people to work together collaboratively in the search for valid, authentic and morally correct and appropriate ways of understanding the world and participating in it (p. 578). Whereas some action research focuses on individuals, participatory action research is communitarian and social, seeking to bring about social change and improvement to the quality of people's lives (Creswell, 2012, pp. 582–3).

McTaggart (1989), Kemmis and McTaggart (2005), Locke *et al*. (2013) and Kemmis *et al*. (2014) suggest several tenets of participatory action research, indicating that it is a social process that focuses on the relationship between individuals and their social environment. It is deliberately practical, seeking to improve social practice and people by having them work on themselves. In doing this it requires authentic participation and is collaborative, establishing self-critical, non-hierarchical communities and partnerships. It is also a recursive and systematic process of learning, with planning, action, analysis and reflection leading to further planning, action, analysis and reflection. This involves

people in theorizing about their own practices and values, testing their own assumptions, values, ideas and practices in real-life practice; in other words it is reflexive, drawing together theory and practice. Participatory action research requires participants to build and keep evidential records of practice, theory and reflection and to provide a reasoned justification to others for their work. These, in turn, require participants to look at and document their own experiences objectively. Participatory action research is part of a political process (e.g. towards democracy) and involves people in making critical analyses of a situation, research and practice. It starts small and in small cycles, with small groups of people, and is critical and emancipatory, with participants addressing and interrogating unjust social structures which limit people's development and self-realization. It also disseminates findings to other practitioners and networks.

For Kemmis *et al*. (2014), adopting a critical/critical theory approach is a *sine qua non* of participatory action research, the major aim of which is to change from practice to praxis (committed, informed, self-realizing action, 'practical philosophy' which links theory, thinking and practice (Carr, 2005)), inspired by the language and operation of possibility, solidarity, open communication and freedom from dominatory and oppressive social conditions.

That there is a coupling of the ideological and political debate here has been brought into focus with the work of Freire (1972) and Torres (1992, p. 56) in Latin America, the latter setting out several principles of participatory action research:

- it commences with explicit social and political intentions that articulate with the dominated and poor classes and groups in society;
- it must involve popular participation in the research process, i.e. it must have a social basis;
- it regards knowledge as an agent of social transformation as a whole, thereby constituting a powerful critique of those views of knowledge (theory) as somehow separate from practice;
- its epistemological base is rooted in critical theory and its critique of the subject/object relations in research;
- it must raise the consciousness of individuals, groups and nations;
- it must lead to transformation and emancipation.

Participatory action research does not mean that all participants need to be doing the same. This recognizes a role for the researcher as facilitator, guide, formulator and summarizer of knowledge, raiser of issues (e.g. the possible consequences of actions, the awareness of structural conditions) (Weiskopf and Laske, 1996, pp. 132–3).

Participatory action research is distinguished not only by its methodology (collective participation) and its outcomes (democracy, voice, emancipation), but by its areas of focus (inequalities of power, grass-roots agendas for change and development, e.g. educational inequality, social exclusion, sexism and racism in education, powerlessness in decision making, student disaffection with a socially reproductive curriculum, elitism in education) and its intention to change society and social situations (cf. Fine, 2010; INCITE, 2010; Kemmis, 2010; Kemmis *et al*., 2014). Importantly here, the agendas and areas of focus are identified by the participants themselves, so they are rooted in reality, are authentic and are 'owned' by the participants and communities themselves.

Participatory action research – people acting and researching on, by, with and for themselves – is a democratic activity (Grundy, 1987, p. 142; David, 2002; Jones and Stanley, 2010; Kemmis *et al*., 2014). This form of democracy is participatory (rather than, for example, representative): a key feature of critical research (and Lewin – a key early figure in action research – advocated democratic workplaces) (Jefferson, 2014, p. 93). It is not merely a form of change theory, but addresses fundamental issues of power and power relationships, for, in according power to participants, action research is an empowering activity (David, 2002; Kemmis *et al*., 2014). Elliott (1991, p. 54) argues that such empowerment has to be at a collective rather than an individual level, as individuals do not operate in isolation from each other, but are shaped by organizational and structural forces.

The issue is important, for it begins to separate action research into different camps (Kemmis, 1997, p. 177). On the one hand are long-time advocates of action research such as Elliott (e.g. 1978, 1991) who, in the tradition of Schwab and Schön, emphasize reflective practice; this is particularly so in curriculum research with notions of the 'teacher-as-researcher' (Stenhouse, 1975) and the reflective practitioner (Schön, 1983, 1987). On the other hand are advocates of the 'critical' action research model, for example, Carr and Kemmis (1986), Kemmis *et al*. (2014).

22.5 Action research as critical praxis

Much of the writing in this type of action research draws on the Frankfurt School of critical theory (see Chapter 3), in particular the early work of Habermas.

(Kemmis (2006) addresses some of the later work of Habermas in his (Habermas's) concern for the 'public sphere'.) Indeed Weiskopf and Laske (1996, p. 123) and Kemmis *et al*. (2014) locate action research, in the German tradition, squarely as a 'critical social science' with an emancipatory interest: to challenge and thence to transform unjust and repressive, alienating social structures. Using Habermas's early writing on knowledge-constitutive interests (1972, 1974), a three-fold typology of action research can be constructed which comprises technical, practical and emancipatory (critical) interests; the classification was set out in Chapter 3, and is the basis for the seminal works of Carr and Kemmis (1986) and Grundy (1987) in the field of action research.

The work of Carr and Kemmis (1986) fuelled a tradition of critical action research and its 'explanatory, normative and practical dimensions' (Hawkins, 2015, p. 466). Critical theory advocates the understanding and ideological interrogation of social conditions, and aims to bring about democracy, equality and social justice, partly by exposing and working on the understandings of participants who are seeking such emancipation and societal transformation and also by looking at the 'conditions of possibility' (p. 466) for such to occur.

Grundy (1987, p. 154) argues that 'technical' action research is designed to render an existing situation more efficient and effective, to improve outcomes of practice (Kemmis, 2009, p. 469). In this respect it is akin to Argyris's notion of 'single-loop learning' (Argyris, 1990), being functional, often short-term and technical. It is also akin to Schön's (1987) notion of 'reflection-in-action' (cf. Luttenberg *et al*., 2016). Elliott (1991, p. 55) suggests that this view is limiting for action research since it is too individualistic and neglects wider curriculum structures, regarding teachers in isolation from wider factors.

By contrast, 'practical' action research is designed to promote teachers' professionalism by drawing on their informed judgement to enable them to act more wisely (Grundy, 1987, p. 154; Kemmis, 2009, p. 470). This underpins the 'teacher-as-researcher' movement, inspired by Stenhouse (cf. Pring, 2015). It is akin to Schön's 'reflection-on-action' and is a hermeneutic activity of understanding and interpreting social situations with a view to their improvement (Luttenberg *et al*., 2016). Echoing this, Kincheloe (2003, p. 42) suggests that action research rejects positivistic views of rationality, objectivity, truth and methodology, preferring hermeneutic understanding (phronesis) (Thomas, 2010) and emancipatory practice. As Kincheloe notes (p. 108), the teacher-as-researcher movement is a political enterprise rather than the accretion of trivial cookbook remedies (a technical exercise). Indeed Luttenberg *et al*. (2016) regard reflection on action as a moral activity, not simply a technical-instrumental matter. Marshall and Rossman (2016) state very clearly (p. 26) that action research eschews the claimed neutrality and objectivity of traditional research in favour of promoting values-based change in their own institutions.

Emancipatory action research has an explicit agenda which is as political as is it educational, promoting social justice (Gibbs *et al*., 2016). Grundy (1987) argues (pp. 146–7) that such emancipatory action research seeks to develop in participants their understandings of illegitimate structural and interpersonal constraints that are preventing the exercise of their autonomy and freedom. These constraints, she argues, are based on illegitimate repression, domination and control. When participants develop a consciousness of these constraints, she suggests, they begin to move from unfreedom and constraint to freedom, autonomy and social justice.

Kincheloe (2003, pp. 138–9) clarifies emancipatory action research as:

- constructing a system of meaning;
- understanding dominant research methods and their effects;
- selecting what to study;
- acquiring a variety of research strategies;
- making sense of information collected;
- gaining awareness of the tacit theories and assumptions which guide practice;
- viewing teaching as an emancipatory, praxis-based act.

'Praxis' here is defined as action informed through reflection, and with emancipation as its goal, a 'morally committed action' (Kemmis, 2009, p. 465), and emancipatory/critical action research is part of a collective and collaborative enterprise to transform social formations, structures and practices that are built into the architecture of our lives and societies and which are deemed to be unsustainable on moral, social, ecological, material, rational, ideological, personal, political and economic grounds (cf. Kemmis, 2009, pp. 470–1, 2010).

Action research, here, empowers individuals and social groups to take control over their lives within a framework of the promotion of rather than the 'suppression of generalizable interests' (Habermas, 1976, p. 113). It commences with a challenge to the illegitimate operation of power and requires participants to question and challenge given value systems. For

Grundy (1987), praxis fuses theory and practice within an egalitarian social order, and action research is designed with a political agenda of improvement towards a more just, egalitarian society. This accords to some extent with Lewin's view that action research leads to equality and cooperation, an end to exploitation and the furtherance of democracy (see also Carr and Kemmis, 1986, p. 163; Jones and Stanley, 2010). Zuber-Skerritt (1996a) suggests that:

> emancipatory action research … is collaborative, critical and self-critical inquiry by practitioners … into a major problem or issue or concern in their own practice. They own the problem and feel responsible and accountable for solving it through teamwork and through following a cyclical process of:
>
> 1 strategic *planning*;
> 2 *action*, i.e. implementing the plan;
> 3 *observation*, evaluation and self-evaluation;
> 4 critical and self-critical *reflection* on the results of points 1–3 and making decisions for the next cycle of action research.
>
> (Zuber-Skerritt, 1996a, p. 3)

Action research, she argues,

> is *emancipatory* when it aims not only at technical and practical improvement and the participants' better understanding, along with transformation and change within the existing boundaries and conditions, but also at changing the system itself or those conditions which impede desired improvement in the system/organization.… There is no hierarchy, but open and 'symmetrical communication'.
>
> (Zuber-Skerritt, 1996a, p. 5)

This form of participatory research forms 'empathetic and compassionate ties' (Hawkins, 2015, p. 468) that hold people together.

The emancipatory interest takes seriously the notion of action researchers as participants in a community of equals. This, in turn, is premised on Habermas's notion of the 'ideal speech situation' (Morrison, 1995a, pp. 99–104; cf. Hawkins, 2015, p. 468). Here action research is construed as reflective practice with a political agenda and in which all participants (and action research is participatory) are equal 'players'. Action research is necessarily dialogical – interpersonal – rather than monological (individual), and communication is an intrinsic element, with communication being among the community of equals (Grundy and Kemmis (1988, p. 87) term this 'symmetrical communication'). Because it is a community of equals, action research is necessarily democratic and promotes democracy. Indeed the search is for consensus (and consensus requires more than one participant), hence it requires collaboration and participation.

The link between Habermas's 'ideal speech situation' and action research is reinforced by Eady *et al*. (2015) in their argument that action research benefits from 'communicative space': a physical, emotional and temporal space in which 'professionals are able to engage in meaningful modes of collaboration, democratic and non-judgmental dialogue' (p. 107), i.e. learning together through dialogue.

Emancipatory action research fulfils the requirements of action research set out earlier by Kemmis and McTaggart (1988, 2005) and Kemmis *et al*. (2014); indeed it could be argued that *only* emancipatory action research (in the threefold typology) has the potential to do this.

Kemmis (1997, p. 177) suggests that the distinction between the two camps (the reflective practitioners and the critical theorists) lies in their interpretation of action research. For the former, action research is an improvement to professional practice at the local, perhaps classroom level, within the capacities of individuals and the situations in which they are working; for the latter, action research is part of a broader agenda of changing education, changing schooling and changing society.

A key term in action research is 'empowerment'; for the former camp, empowerment is largely a matter of the professional sphere of operations, achieving professional autonomy through professional development. For the latter, empowerment concerns taking control over one's life within a just, egalitarian, democratic society. Whether the latter is realizable or utopian is a matter of critique of this view. Where is the evidence that critical action research either empowers groups or alters the macro-structures of society? Is critical action research socially transformative?

Several concerns have been levelled at emancipatory action research (Gibson, 1985; Morrison, 1995a, 1995b; Somekh, 1995; Melrose, 1996; Grundy, 1996; Weiskopf and Laske, 1996; Webb, 1996; McTaggart, 1996; Kemmis, 1997; Elliott, 2005; Hadfield, 2012), including the views that:

1. it is utopian and unrealizable;
2. it is too controlling and prescriptive, seeking to capture and contain action research within a particular mould – it moves towards conformity;
3. it adopts a narrow and particularistic view of emancipation and action research, and how to undertake the latter;

4 it undermines the significance of the individual teacher-as-researcher in favour of self-critical communities. (Kemmis and McTaggart (1992, p. 152) pose the question 'why *must* action research consist of a *group* process?');
5 the threefold typification of action research is untenable;
6 it assumes that rational consensus is achievable, that rational debate will empower all participants (i.e. it understates the issue of power, wherein the most informed are already the most powerful. Grundy (1996, p. 111) argues that the better argument derives from the one with the most evidence and reasons, and that these are more available to the powerful, thereby rendering the conditions of equality suspect);
7 it overstates the desirability of consensus-oriented research (which neglects the complexity of power);
8 power cannot be dispersed or rearranged simply by rationality;
9 action research as critical theory reduces its practical impact and confines it to the commodification of knowledge in the academy;
10 it will promote conformity through slavish adherence to its orthodoxies;
11 is naive in its understanding of groups and celebrates groups over individuals, particularly the 'in-groups' rather than the 'out-groups';
12 privileges its own view of science (rejecting objectivity) and lacks modesty;
13 privileges the authority and supremacy of critical action research over other equally positive forms of action research;
14 critical action research has framed rather than changed or shaped social praxis;
15 is elitist whilst purporting to serve egalitarianism;
16 assumes an undifferentiated view of action research;
17 is attempting to colonize and redirect action research.

This critique serves to remind the reader that critical action research is problematical. It may be just as controlling as those controlling agendas that it seeks to attack (Morrison, 1995b). Indeed Melrose (1996, p. 52) suggests that because critical research is itself value-laden, it abandons neutrality; it has an explicit social agenda that, under the guise of examining values, ethics, morals and politics which operate in a particular situation, is actually aimed at transforming the status quo. This echoes the critique of non-neutral research by Hammersley (2000, 2014) in Chapter 3.

For a simple introductory exercise for understanding action research, see the accompanying website.

22.6 Action research and complexity theory

Action research links with participatory research and has affinities with complexity theory. Phelps and Graham (2010, p. 184) argue that action research 'can readily accommodate the key tenets of complexity theory' and that there is a 'deep complementarity' between them. For example, they note (p. 187) that action research accepts that systems are unpredictable, open and non-linear. It resonates with issues of adaptation to environment and can lead to bifurcation, i.e. when a system moves from one 'point of stability to another' (p. 190). It celebrates the interaction of participants and requires both feedback and feed forward. It is reflective and shows an interest in 'exceptions' or outliers (which can lead to major change) (p. 194). It is not concerned with controlling variables, and accepts that the systems in which it takes place are complex and dynamic (see also Davis and Sumara, 2005, p. 455). Action research, like complexity theory, celebrates self-organization, with new states and situations emerging from tipping points (Morrison, 2008): self-organized criticality (Bak, 1996). Similarly, Luttenberg *et al*. (2016) note that action research, and the reflection that it involves, produces outcomes and processes that are open, dynamic, non-linear, adaptive and co-adaptive (between components), and emergent, and that action research itself can be regarded as a complex system (pp. 6–12).

22.7 Procedures for action research

There are several ways in which steps in action research have been analysed. Lewin (1946, 1948) codified the action research process into four main stages: planning, acting, observing and reflecting (cf. New South Wales (NSW) Department of Education and Training, 2010). This operates in a cyclical process, with one cycle of this four-step approach leading into the subsequent four-step cycle. The NSW Department of Education and Training (2010) suggest that: between 'planning' and 'acting' come 'identifying', 'informing' and 'organising'; between 'acting' and 'observing' come 'trialling', 'collecting' and 'questioning'; between 'observing' and 'reflecting' come 'analysing', 'reporting' and 'sharing'; and between 'reflecting' and the new cycle comes 'planning', 'evaluating', 'implementing' and 'revising' (p. 3). Piggot-Irvine *et al*. (2015) note that, in fact, action research has no clearly defined ending (p. 549) and the end of one cycle leads into the beginning of the next.

Lewin (1946, 1948) suggests that action research commences with a general idea and data are sought about the presenting situation. The successful outcome

of this examination is the production of a plan of action to reach an identified objective, together with a decision on the first steps to be taken. Lewin acknowledges that this might involve modifying the original plan or idea. The next stage of implementation is accompanied by ongoing fact-finding to monitor and evaluate the intervention, i.e. to act as a formative evaluation. This feeds forward into a revised plan and set of procedures for implementation, themselves accompanied by monitoring and evaluation. Lewin (1948, p. 205) suggests that such 'rational social management' can be conceived of as a spiral of planning, action and fact-finding about the outcomes of the actions taken.

McKernan (1991, p. 17) suggests that Lewin's model of action research is a series of spirals, each of which incorporates a cycle of analysis, reconnaissance, reconceptualization of the problem, planning of the intervention, implementation of the plan and evaluation of the effectiveness of the intervention. Ebbutt (1985) adds that feedback within and between each cycle is important, facilitating reflection. This is reinforced in the model of action research by Altricher and Gstettner (1993) where, though they have four steps (p. 343) – (a) finding a starting point, (b) clarifying the situation, (c) developing action strategies and putting them into practice, (d) making teachers' knowledge public – they suggest that steps (b) and (c) need not be sequential, thereby avoiding the artificial divide that might exist between data collection, analysis and interpretation.

Zuber-Skerritt (1996b, p. 84) sets emancipatory (critical) action research into a cyclical process of: '(1) strategic planning, (2) implementing the plan (action), (3) observation, evaluation and self-evaluation, (4) critical and self-critical reflection on the results of (1) – (3) and making decisions for the next cycle of research.'

Bassey (1998) sets out eight stages in action research:

Stage 1: Defining the inquiry.
Stage 2: Describing the educational context and situation.
Stage 3: Collecting evaluative data and analysing them.
Stage 4: Reviewing the data and looking for contradictions.
Stage 5: Tackling a contradiction by introducing change.
Stage 6: Monitoring the change.
Stage 7: Analysing evaluative data about the change.
Stage 8: Reviewing the change and deciding what to do next.

Moroni (2011), deliberately echoing Zuber-Skerritt (1996b), sets out a five-step process of action research:

1. *Diagnosis* of a problem, which involves what is to be investigated and the purposes of the action research, for example, to answer a research question, to test a hypothesis, to improve practice.
2. *Planning* an intervention to address the problem, which involves considering what the intervention will comprise, what data are required and how to gather them, and what data-collection instruments are required.
3. *Action:* putting the intervention into practice, which involves consideration of timing and duration, participants, contents of the intervention.
4. *Assessment*: how far the intervention has met its objectives of solving the problem, which involves consideration of how to analyse and interpret the data.
5. *Critical reflection and communication of learning*: reflection on the experience and what has been learned, and sharing this, which involves consideration of how to disseminate the findings.

McAteer (2013) sets out a five-stage process of action research: identifying the research question; finding out about the present situation; identifying changes ('action steps') that can be made; evaluating the effects of such changes; and revising the original question as a consequence of the findings of the research (pp. 32–3).

An alternative, eight-stage model is thus:

Stage 1: Decide and agree one common problem that you are experiencing or need that must be addressed.
Stage 2: Identify some causes of the problem (need).
Stage 3: Brainstorm a range of possible practical solutions to the problem, to address the real problem and the real cause(s).
Stage 4: From the range of possible practical solutions decide *one* of the solutions to the problems, perhaps what you consider to be the most suitable or best solution to the problem. Plan how to put the solution into practice.
Stage 5: Identify some 'success criteria' by which you will be able to judge whether the solution has worked to solve the problem, i.e. how will you know whether the proposed solution, when it is put into practice, has been successful. Identify some practical criteria which will tell you how successful the project has been.
Stage 6: Put the plan into action; monitor, adjust and evaluate what is taking place.
Stage 7: Evaluate the outcome to see how well it has addressed and solved the problem or need, using the success criteria identified in Stage 5.
Stage 8: Review and plan what needs to be done in light of the evaluation.

The key features of action research here are:

- it works on, and tries to solve, real, practitioner-identified problems of everyday practice;
- it is collaborative and builds in teacher involvement;
- it seeks causes and tries to work on those causes;
- the solutions are suggested by the practitioners involved;
- it involves a divergent phase and a convergent phase;
- it plans an intervention by the practitioners themselves;
- it implements the intervention;
- it evaluates the success of the intervention in solving the identified problem.

We set out below our own eight-step process of action research, drawing together the several strands and steps of action research.

Step 1

The first step involves the identification, evaluation and formulation of the problem (widely defined, e.g. to include a need for innovation) perceived as critical in an everyday teaching situation. McAteer (2013) suggests that the problem should: be related to improving one's own practice; enable explanations and hypotheses to be developed (relating them to a broader base of theory); be within the action researcher's own power and control to change; and be professionally and personally important and pertinent (p. 28).

Step 2

The second step involves preliminary discussion and negotiations among the interested parties – teachers, researchers, advisers, sponsors, possibly – which may culminate in a draft proposal. This may include a statement of the questions to be answered (e.g. 'Under what conditions can curriculum change be best effected?' 'What are the limiting factors in bringing about effective pedagogical change?' 'What strong points of action research can be employed to bring about assessment change?'). The researchers in their capacity as consultants (or sometimes as programme initiators) may draw upon their expertise to bring the problem more into focus, possibly determining causal factors or recommending alternative lines of approach to established ones. This is often the crucial stage for the venture as it is at this point that the seeds of success or failure are planted, for, generally speaking, unless the objectives, purposes and assumptions are made perfectly clear to all concerned, and unless the role of key concepts is stressed, the enterprise can easily miscarry.

Step 3

The third step, in some circumstances, may involve a review of the research literature to find out what can be learned from comparable studies, their objectives, procedures and problems encountered.

Step 4

The fourth step may involve a modification or redefinition of the initial statement of the problem at Step 1. It may now emerge in the form of a testable hypothesis, or as a set of guiding objectives. Sometimes change agents deliberately decide against the use of objectives on the grounds that they have a constraining effect on the process itself. It is also at this stage that assumptions underlying the project are made explicit (e.g. in order to effect curriculum changes, the attitudes, values, skills and objectives of the teachers involved must be changed).

Step 5

The fifth step is concerned with the selection of research procedures – sampling, administration, choice of materials, methods of teaching and learning, allocation of resources and tasks, deployment of staff and so on. Here it must be stated that embedded within the overall scope of the term 'action research' might be a number of different research designs that include different methods of gathering data. A piece of action research might include, for example:

- an initial and end-of-intervention survey (a pre- and post-survey);
- an experimental or quasi-experimental design (e.g. where some students/teachers are involved in the intervention and some are not, or where pre- and post-testing of students/teachers is undertaken);
- a longitudinal study (over the duration of the intervention);
- participant and non-participant observation;
- interviews and field notes;
- one or more case studies;
- documentation from, and about, participants;
- questionnaire data.

In this respect readers are advised to go to the chapters in this book that address these methods, in particular on case study, experiments and quasi-experiments, and observation. Many novice researchers are unsure whether their research is action research or a case study; indeed it may be both, but a distinguishing feature may be whether the research involves an intervention on the part of the researcher(s), or whether the data are largely only collected. If it is the former

– concerning change, development and intervention – then it may be action research, whereas if it is largely the latter, then it may be more of a case study; one has to be very cautious in making this distinction because there can be gross overlaps between the two.

As action research is intended to bring about a change, with an intervention involved, then the researcher may wish to use an experimental or quasi-experimental approach in the action research in an attempt to identify causality through a controlled intervention, with control and experimental groups (see Chapter 20).

Step 6

The sixth step is concerned with the choice of the evaluation procedures to be used and takes into consideration that evaluation in this context will be continuous.

Step 7

The seventh step is the implementation of the intervention itself (over varying periods of time). It includes: the conditions and methods of data collection (e.g. fortnightly meetings, the keeping of records, interim reports, final reports, the submission of self-evaluation and group-evaluation reports, etc.); the monitoring of tasks and the transmission of feedback to the research team; and the classification and analysis of data.

Step 8

The eighth step involves the interpretation of the data; inferences to be drawn; and overall evaluation of the project (cf. Woods, 1989). Discussions of the findings take place in the light of previously agreed evaluative criteria. Errors, mistakes and problems are considered. A general summing-up may follow this in which the outcomes of the project are reviewed, recommendations made, and arrangements for dissemination of results to interested parties decided.

At every stage there is reflection and self-reflection, addressing reflexivity (discussed below). This eight-step process is set out in Figure 22.1. It does not *necessarily* follow a linear sequence, and steps may be recursive and in a different sequence. As Figure 22.1 indicates, evaluation and reflection accompany every stage of the process. Reflection can be descriptive (personal, looking back at what has happened), perceptive (e.g. emotional), receptive (relating views of others to one's own views), interactive (lining the past and present to future action) and critical (interrogating the context in which the teacher operates) (McAteer, 2013, p. 26). It can engage a retrospective analysis of critical incidents: those which make a significant difference or sudden solution to a situation (p. 72).

This is a basic framework; much activity of an incidental and possibly ad hoc nature will take place in and around it. This may comprise discussions among teachers, researchers and students; regular meetings among teachers or schools to discuss progress and problems and to exchange information; possibly regional conferences; and related activities, all enhanced by current hardware and software.

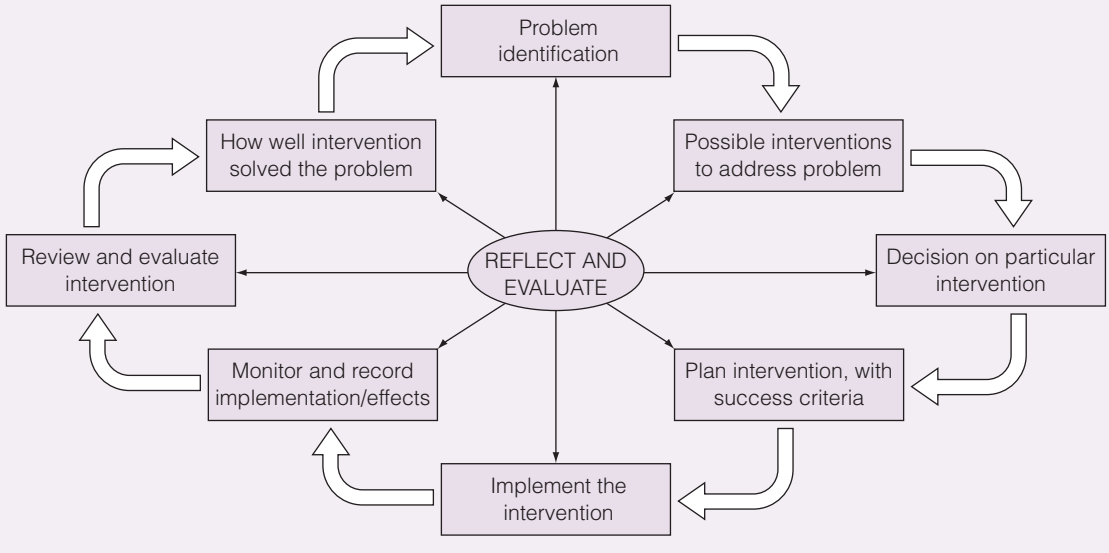


FIGURE 22.1 A framework for action research

Hopkins (1985), McNiff *et al*. (1996), McNiff and Whitehead (2009) and McNiff (2010) offer much practical advice on the conduct of action research, including 'getting started', operationalization, planning, monitoring and documenting the intervention, collecting data and making sense of them, using case studies, evaluating the action research, ethical issues and reporting. We urge readers to go to these helpful sources. These are useful introductory sources and guides for practice, in particular McNiff (2010). Indeed McNiff (2010) takes the reader, novice or experienced, through key features of action research, including: what it is, how it differs 'traditional research' and how it fits western research traditions; why people should do it; how to do it; who can do it; where to do it; what it involves; how to start and how to identify a concern; matters of values; kinds of action; how to reflect on its different aspects and elements; action planning, data collection and analysis; how to ensure that conclusions are fair, valid and reliable; how to judge its significance; implications of the action research for different parties; how to write up, report and disseminate action research, and how to use it in the development of a professional portfolio. Along the way, she raises a range of clearly expressed questions, points for reflection, including action perspectives and research perspectives (p. 22), and 'difficult questions' such as 'Whose practice?' 'Whose research?' 'Whose voice?' 'Whose theory?' and 'Who speaks?' (p. 57). She stresses the importance of setting evaluative criteria. Without success criteria it is impossible for the researcher to know whether, or how far, the action research has been successful. Action researchers could ask themselves 'How will we know whether we have been successful?' Her volume is a tour de force for action researchers.

Kemmis and McTaggart (1992, pp. 25–7) offer a useful series of observations for beginning action research which involves:

- convening and organizing a group of action researchers as participants, even if the group is small;
- being prepared to start small and expanding over time, keeping a focus on the longer term and larger issues (e.g. whole-school issues) as well as the shorter term and immediate issues;
- setting time frames and actions in them, including support and development activity;
- building in tolerance, involvement of all participants and support as participants learn by doing, reflecting on what is happening and taking responsibility for actions and consequences;
- scrupulously recording developments and progress in a timely fashion, and disseminating these beyond the group of action researchers, indicating clearly the progress that has been made;
- bring in outsiders (e.g. external consultants) where appropriate, for example, to provide legitimation for the action research;
- ensuring that the action research enables participants to put their educational values into practice.

It is clear from this list that action research is a blend of practical and theoretical concerns; it is both action and research.

In conducting action research the participants can be both methodologically eclectic and can use a variety of instruments for data collection: questionnaires, diaries, interviews, case studies, observational data, experimental design, field notes, photography, audio and video recording, sociometry, rating scales, biographies and accounts, documents and records, in short the full *gamut* of techniques (for a discussion of these see Parts 3 and 4 of the present volume).

22.8 Reporting action research

McNiff and Whitehead (2009, p. 15) suggest that, in reporting action research, it is important to note not only the action but the research element, including the rigorous methodology and interpretation of data, and that it is important to state clearly:

- the research issue and how it came to become a research issue in the improvement of practice;
- the methodology of, and justification for, the intervention, and how it was selected from among other possible interventions;
- how the intervention derived from an understanding of the situation;
- what data were collected, when and from whom;
- how data were collected, processed and analysed;
- how the ongoing intervention was monitored and reviewed;
- how reflexivity was addressed;
- what were the standards and criteria for success, and how these criteria were derived;
- how conclusions were reached and how these were validated;
- what and how the researcher learnt as a consequence of the action research;
- how practice was changed as a consequence of the findings.

The authors note that validity is a key concern in reporting, that warrants have to be justified for the conclusions drawn, that these warrants reside in the

evidential trail provided in the research (p. 23) and that reflection and reflexivity must be demonstrated (p. 28).

It was noted at the start of this chapter that the goal of action research is improvement; therefore the report must indicate not only what the improvement was, but that it was attributable to the intervention and not to other factors, i.e. that causality is demonstrated. This requires a level of rigour that is indicated in the 'research' part of the 'action research'. More than this, given that action research concerns research, the report should indicate not only how the research led to improvement in practice, but how the action research in question contributes to the expansion of knowledge, scholarship and scholarly enquiry, i.e. what significance the research has for both the academic and professional communities. The report, then, serves a dual set of criteria: (a) criteria for the planning, conduct, reporting and evaluation of the research; and (b) criteria for the planning, conducting, reporting and evaluation of practice/action.

Given that the intervention is into a 'real-life' situation, it is important to include in the report some information about the 'real-life' context of the intervention, so that the reader has a clear picture of this. This means that the report must include necessary descriptive data (McNiff and Whitehead, 2009, p. 37), together with scholarly enquiry (e.g. a literature review), explanations, reflections, research methodology, data collection, analysis and interpretation, consideration of alternative explanations, and, of course, evidence that there has been an improvement in practice and in the development of the researcher (e.g. in terms of pedagogy, subject knowledge, researcher ability and skills, reflective capacity).

Action research may be reported in narrative form (e.g. McNiff and Whitehead, 2009, p. 49; McNiff, 2010) and must be written with the reader in mind. An action research report should address (cf. McNiff and Whitehead, 2009, p. 56; McNiff, 2010):

- the action researcher's concern and the reason for that concern;
- an indication of the presenting situation at the start of the action research;
- a review of what, how and why the action researcher moved into action and reflection;
- what methodology, design and data were used in the action research (e.g. it was suggested earlier that embedded in action research might be a case study, an experimental or quasi-experimental approach, a survey, an ethnography);
- what were the research questions;
- what were the problems that the action research was intended to address/solve;
- what possible interventions were considered, and why some of these were rejected/accepted (e.g. on what criteria);
- how the intervention was planned and implemented;
- how ongoing data were gathered, processed and used during the action research;
- what were the roles of the action researcher;
- what was discovered during, and as a consequence of, the action research;
- what conclusions were drawn, and how they were valid (their warrants);
- an indication of the significance of the action research – for action and for research;
- an indication of how practice was modified and improved as a consequence of the action research;
- an indication of, and justification for, the success criteria used to evaluate the action research;
- the reflections of the action researcher, together with evidence of growth in reflective ability (and the criteria used to evaluate this).

The action researcher has to adopt a potentially schizophrenic stance to the action and the research, being both in it and of it, but also having to stand back from the situation and viewing it with as much objectivity as possible; subjectivity and objectivity (or, perhaps better, relative subjectivity and objectivity) are combined in a single action researcher.

22.9 Reflexivity in action research

The analysis so far has made much of the issue of reflection, be it reflection-in-action, reflection-on-action or critical reflection. Reflection, it has been argued, occurs at every stage of action research. Beyond reflection, *reflexivity* is central to action research, because the researchers are also the participants and practitioners in the action research – they are part of the social world that they are studying (Hammersley and Atkinson, 1983, p. 14). Hall (1996, p. 29) suggests that reflexivity is an integral element and epistemological basis of emancipatory action research because it takes as its basis the view of the construction of knowledge in which: (a) data are authentic and reflect the experiences of all participants; and (b) democratic relations exist between all participants in the research; the researcher's views (which may be theory-laden) do not hold precedence over the views of participants.

Reflexivity requires a self-conscious awareness of the effects that the participants-as-practitioners-and-researchers are having on the research process, how their values, attitudes, perceptions, opinions, actions, feelings etc. are feeding into the situation being studied

(akin, perhaps, to the notion of counter-transference in counselling). The participants-as-practitioners-and-researchers need to apply to themselves the same critical scrutiny that they are applying to others and to the research, as discussed in Chapter 14.

Reflexivity also links to awareness of possible bias, in that the practitioner is also the researcher and may not be entirely disinterested. For example, in an attempt to impress a senior manager, a teacher who is an action researcher may present a rosier picture of the outcome of the action research than is really the case, or, by contrast, a teacher who may be pressing for increased resources may present the outcome more negatively than it is. Here ethics, validity and political agendas coincide.

22.10 Ethical issues in action research

Action research is not exempted from the ethical issues that were identified in Chapter 7. It requires the informed consent of participants, options for teachers/students not to take part, and with no penalty (Nolen and Vander Putten, 2007), and confidentiality and autonomy of participants to be respected. Whilst referring the reader to Chapter 7, we also note that there is a blurred dividing line between the teacher qua teacher and qua researcher, and that effective teaching also concerns effective researching. Perhaps, also, the fact that minors attend school on a compulsory basis already gives the teacher automatic right to research them as part of her everyday teaching. Where is the dividing line?

Gibbs *et al*. (2016) draw attention to the ethical challenges faced by insider and outsider action researchers, for example: 'fiduciary responsibilities' to whom and for what; the principle of 'do no harm' (p. 7); ownership of intellectual property (pp. 11–12).

Locke *et al*. (2013, pp. 109, 119–20) and Denscombe (2014, p. 127) identify a range of ethical issues that action research should address:

- how to maintain confidentiality whilst acknowledging others' contributions, and how to address the balance between confidentiality and disclosure;
- the potential knock-on effects of the action research on participants and other relevant parties;
- how to avoid doing harm to participants (e.g. from disclosure);
- how to corroborate the data and interpretation;
- the need to seek approval and clearance for the research (i.e. simply because it is action research does not exempt the researcher from seeking ethical approval);
- how to address 'bad news', i.e. reporting negative results and presenting results in a bad light;
- informed consent;
- the increased workload on participants that is likely to come with action research;
- protection of vulnerable people;
- recognition that protection from harm trumps personal beneficence or benefit;
- equitable selection and inclusion of participants.

Locke *et al*. (2013) set out key ethical principles for action research (pp. 113–14):

- respect for all participants as stakeholders who genuinely share decisions (the 'principle of inclusivity');
- respect for all participants, in whatever roles, as 'full members' of the action research group (the 'principle of maximal participant recognition');
- aims, content and operation of the research and ownership of the data and report agreed and decided in consultation by all participants ('the principle of negotiation and consensus');
- rights of withdrawal and renegotiation of grounds for participation ('the principle of communicative freedom');
- use of plain, comprehensible language by all parties ('the principle of plain speaking');
- ensuring that all 'members' collaboratively adjudicate the moral rightness of the aims, processes and understandings of the research ('the principle of right action');
- ensuring questioning of, and transparency in, the 'discursive assumptions' that participants bring to the research ('the principle of critical self-reflexivity');
- ensuring that the feelings of all participants are respected and count ('the affective principle').

22.11 Some practical and theoretical matters

Much has been made in this chapter of the democratic principles that underpin some types of action research. The ramifications of this are several. For example, there must be a free flow of information between participants and communication must be extensive (Elliott, 1978, p. 356). Further, communication must be open, unconstrained and unconstraining – the force of the better argument in Habermas's 'ideal speech situation'. That this might be problematic in some organizations has been noted by Holly (1984, p. 100), as action research and schools are often structured differently, with schools being hierarchical, formal and bureaucratic whilst action research is collegial, informal, open,

collaborative and crosses formal boundaries. In turn this suggests that, for action research to be successful, the conditions of collegiality have to be present, echoing Habermas's 'ideal speech situation', for example (Morrison, 1995a, 1998, pp. 157–8, 2011, p. 153):

- participatory approaches to decision making;
- democratic and consensual decision making;
- shared values, beliefs and goals;
- equal rights to determine policy;
- equal voting rights on decisions;
- the deployment of sub-groups who are accountable to the whole group;
- shared responsibility and open accountability;
- an extended view of expertise;
- judgements and decisions based on the power of the argument rather than the positional power of the advocates;
- orientation to a common interest ascertained without deception;
- everyone's freedom to enter a discourse, to check questionable claims, to evaluate explanations, to modify a given conceptual framework and to reflect on the nature of both knowledge and political will;
- everyone's freedom to assess justifications and to alter norms;
- mutual understanding between participants;
- recognition of the legitimacy of each subject to participate in the dialogue as an autonomous and equal partner, with equal opportunity for discussion;
- discussion to be free from domination and distorting or deforming influences;
- all motives except for the cooperative search for truth are excluded;
- the speech act validity claims of truth, legitimacy, sincerity and comprehensibility are all embodied;
- illocutions (where the outcome is open or negotiable) replace perlocutions (achieving a given, predetermined, non-negotiable outcome by saying something);
- shared ownership of decisions and practices.

Zuber-Skerritt (1996b, p. 90) suggests that the main barriers to emancipatory action research are: (a) single-loop learning (rather than double-loop learning) (Argyris, 1990); (b) overdependence on experts or seniors to the extent that independent thought and expression are stifled; (c) an orientation to efficiency rather than to research and development (one could add here 'rather than to reflection and problem posing'); (d) a preoccupation with operational rather than strategic thinking and practice. She suggests (1996a, p. 17) four practical problems that action researchers might face:

- How can we formulate a method of work which is sufficiently economical as regards the amount of data gathering and data processing for a practitioner to undertake it alongside a normal workload, over a limited timescale?
- How can action research techniques be sufficiently specific to enable a small-scale investigation by a practitioner to lead to genuinely new insights, and avoid being accused of being either too minimal to be valid, or too elaborate to be feasible?
- How can these methods, given the above, be readily available and accessible to anyone who wishes to practise them, building on the competencies which practitioners already possess?
- How can these methods contribute a genuine improvement of understanding and skill, beyond prior competence, in return for the time and energy expended – that is, a more rigorous process than that which characterizes positivist research?

Another issue of some consequence concerns headteachers' and teachers' attitudes to the possibility of change as a result of action research. Hutchinson and Whitehouse (1986), for example, note possible resistance from headteachers and teachers themselves.

Further, Jones and Stanley (2010) comment that action research involving university researchers and public stakeholders seriously challenges 'the democratic principles commonly associated with this genre of critical enquiry' (p. 161), as micro-politics can frustrate the endeavour to be truly democratic. Kemmis (2006) and Gibbs *et al*. (2016) question whether individual or even collaborative action research can really live up to its claim to radically challenge and change injustice, i.e. whether its putative emancipatory potential is realizable in practice, and that it is more descriptive of the reflective process than being evaluative or emancipatory (p. 7).

Piggot-Irvine *et al*. (2015) comment that more needs to be done to evaluate the outcomes and impact of action research rather than its predominant focus on process, and they indicate criteria, foci and indicators for evaluating precursors/foundations, processes/activities, outcomes and impacts. Similarly, Heikkinen *et al*. (2012) suggest five principles for the validation of action research: *historical continuity* (locating the research in its antecedents and unfolding course of action); *reflexivity* (awareness and disclosure of the impact of personal experience on the research); *dialectics* (inclusion of multiple voices in the research and its interpretation); *workability and ethics* (whether the research has led to changes in practice and addresses ethical issues transparently); and *evocativeness* (how

effectively the research evokes images and emotions in the reader). Pring (2015, p. 160) adds that including externality in action research can increase the public perception of its validity and rigour, though how far this addresses 'objectivity' is a moot point.

22.12 Conclusion

Action research is a potential means of empowering teachers, though this chapter has questioned the extent of this. As a research device it combines six notions:

1. A straightforward cycle of identifying a problem, planning an intervention, implementing the intervention, evaluating the outcome.
2. Reflective practice.
3. Political emancipation.
4. Critical theory.
5. Professional development.
6. Participatory practitioner research.

Action research is a flexible, situationally responsive methodology that offers rigour, authenticity and voice. This chapter has tried to expose both the attractions and problematic areas of action research. In its thrust towards integrating action and research one has to question whether this is an optimistic way of ensuring that research impacts on practice for improvement, or whether it is a recessive hybrid. There are several journals that focus on action research in education, for example: *Educational Action Research* and *Action Research Reflective Practice*, and some key websites:

http://cadres.pepperdine.edu/ccar/about.html (the Centre for Collaborative Action Research)
www.jeanmcniff.com (the website of Jean McNiff)
www.actionresearch.net (the website of Jack Whitehead)
http://aral.com.au (the website of Bob Dick)

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Virtual worlds, social network software and netography in educational research

CHAPTER 23

Stewart Martin

Virtual worlds, social network software and netography are key features of contemporary concern in educational research. This chapter introduces researchers to this field, addressing:

- key features of virtual worlds
- social network software
- using virtual worlds and social media in educational research
- netography, virtual worlds and social media network software
- opportunities for research with virtual worlds, social network software and netography
- ethics
- guidelines for practice
- data

23.1 Introduction

Virtual worlds are computer-based multi-user simulations that are shared online by individuals, who appear as a graphical character: an interactive icon or *avatar*. Although creative inventions, these environments have *persistence*; they continue to exist and change with possible consequences for an avatar, even when a user is not present online, because of the ongoing actions of other online users. Current computer technology enables virtual worlds to offer increasingly convincing three-dimensional imagined environments for individuals to explore alone or as a member of a community and, because of their great versatility and realistic appearance, they are commonly dedicated to entertainment or social interaction or are found in military or commercial applications. However, they are increasingly used in educational and research settings for exploring languages, ethics, management, history, psychology, design, pedagogy and a rapidly widening range of other areas (see Peachey *et al*., 2010; Hinrichs and Wankel, 2011; Hunsinger and Krotoski, 2012; Gregory *et al*., 2015).

In a virtual world, individuals can project their actions, views or values through their avatar and receive feedback from others in the system.[1] Projection techniques can encourage a sense of authenticity to externalize the self and create an impression of presence in an environment that we are not physically part of. This relies on the 'willing suspension of disbelief" (Coleridge, 1817) to create an 'illusory shift in point of view' (Dennett, 1978, p. 312) as well as on the use of our own knowledge, imagination and enthusiasm (Zhao, 2003). Our sense of being present in the physical world appears to be an essential component of consciousness but is not something we normally think about unless prompted by a displacement of our self-perception, for example, through a dream, literature, film or theatrical experience. Any relative lack of realism is not therefore an obstacle to user acceptance of a virtual world and many early examples were effective and engaging, despite having poor graphics; some included no pictorial images at all (Nelson and Erlandson, 2012; Martin, S., 2014).

23.2 Key features of virtual worlds

Virtual worlds are often three-dimensional, visually realistic and attractively designed; they can be useful in stimulating participants' imagination where the sense of being in a real place may be important, for example, to encourage engagement in exploring moral dilemmas (Martin, 2015). Creating a sense of life as a virtual experience may also be valuable for investigating socially sensitive issues by invoking a feeling of a 'safe distance' between the individual and the thing being explored, as the avatar becomes the presenter of a particular perspective or behaviour. Such 'projection' may help to displace any associated difficulty or stress away from the individual and ease the exploration of sensitive or highly charged issues (Freud, 1936). Virtual worlds can also be valuable for developing otherwise inaccessible or impossible environments to explore

human interaction, agency, values and perceptions, such as historical locations and experiences, future imagined scenarios or uninhabitable settings.

These features make virtual worlds particularly effective training environments, where the depiction of a hostile environment (e.g. fire-fighter training for casualty location) can be enhanced to offer more realism without physical risk, or otherwise impossible experiences can be created, such as that of being immobile or disabled, or being a member of a different social or ethnic group.

The low risk and 'repeatability' of experiences in virtual worlds have advantages not just because of the safety in otherwise dangerous or unpredictable environments, but also because, despite their heightened sense of realism compared to other approaches, they afford *discardable experiences* at relatively lower personal, experiential or emotional cost, for example: when training armed forces in decision making in pressured situations; or training medics to treat 'real' casualties; or advancing views, proposals or identity depictions in hostile situations (Waller *et al*., 1998; Martin, S., 2013). Virtual worlds therefore have advantages of *economy* (being cheaper to create and use than real-life settings); of *visibility* (important things can be made clearer and more accessible); of *control* (much more control of the setting is possible than in real life); and of *safety*, where situations can be used that in real life would be too dangerous, difficult, sensitive or ethically questionable (Bailey, 2007; Nelson and Erlandson, 2012).

Together with a configurable avatar, these features make virtual worlds effective technologies for studying and affording rehabilitation experiences for individuals who have undergone traumatic experiences such as domestic violence, or by offering opportunities for people to make otherwise unavailable choices regarding actions, gender or personality. Their potential for exploiting role/real playing and the blending of physical real-world activity with virtual activity offers scope for research using innumerable scenarios. These are environments in which participants could project, share and reflect on their own and others' actions and views, which may lead to the surfacing and exploration of further, often sensitive, issues for investigation and possible resolution, the reconciliation of potential conflict and disagreement, and hence to the development and enlargement of understanding. Although the Internet and its increasingly varied digital spaces offer some distortions in the portrayal of everyday life, through the creation of virtual spaces and communication within them, participants can be encouraged to be more open, honest and authentic in disclosing their views, their values and their beliefs about real or created situations and issues.

23.3 Social network software

Social network software is constantly developing, increasingly multimodal and dynamic, and is used for a wide range of purposes including communication, self-expression, maintaining friendships, sharing information and for enjoyment. Boyd and Ellison (2007) define them as 'web-based services that allow individuals to construct a public or semi-public profile within a bounded system' (p. 211); Kaplan and Haenlein (2010) add that they are also 'a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content' (p. 61). All media foster communication and so to some degree are social by definition (Papacharissi, 2011) and it may often be useful to use one or more of them as part of data gathering and communication in research.

Social media vary and include networking services for sharing personal information (e.g. Facebook); sites for content sharing or discussion (e.g. Flickr, YouTube, Instagram, WhatsApp, Digg or Reddit); forums (e.g. wikis such as wiktionary.org); professional networking sites (e.g. LinkedIn); blogs (e.g. Twitter, Blogspot); virtual social worlds (e.g. Habbo Hotel, Club Penguin, Second Life); and massive multiplayer online role-playing games (MMORPGs) such as World of Warcraft, Final Fantasy and Skyrim. Although used for different purposes, these technologies can have features in common; platforms such as Twitter or Facebook have shared content that is generated, blended or re-posted by users.

23.4 Using virtual worlds and social media in educational research

The popularity of social network sites reflects the significance of communication in contemporary society and influences values, ideas and behaviours that are often of interest to researchers. Digital spaces have architectures that affect the way people work and live in them, and social media offer opportunities to circumvent traditional power relationships and create a greater likelihood that individuals may feel free to express themselves. Virtual worlds and social media can therefore help individuals to communicate more freely and engage with a kind of 'ideal speech' (Habermas, 1979, 1982, 1984, 1987b; see Chapter 14 of the present volume). These ideal speech features can support research by encouraging participants to more openly:

- enter a discourse and check questionable claims;
- evaluate explanations and assess justifications;

- modify a given conceptual framework and influence norms;
- reflect on the nature of political views and action;
- develop mutual understanding between participants, accept that consensus is derived from the better argument and not from the influence or authority of individual participants, and accept that the cooperative search for truth should be the only motive;
- recognize the legitimacy of others to participate in the dialogue as an autonomous and equal partner, and to promote equal opportunity for discussion.

(after Morrison, 1995a, p. 102)

These features may influence who participates and what kinds of things are shared, as not everyone may feel comfortable in such contexts; this should be borne in mind when considering the topic and desired participants for a research project. Cultural or individual differences may also predispose some individuals to accept or decline an invitation to participate and so affect the sample, with implications for the representativeness of a study and for the nature and robustness of its possible conclusions.

Since its founding in 2006 as a simple way to help friends stay in touch, the microblogging and social network site Twitter has become a popular communication medium alongside its earlier counterpart Facebook. The use of Twitter and Facebook in many western high schools is common as young people create content and tag and communicate this to interact socially in school and in informal spaces beyond school. Users typically set Twitter accounts to 'public' to share their social lives widely and this information can be used to study connections between people, their self-expression and engagement with resources and learning communities. Both Facebook and Twitter, as 'always on' technologies, are thought to particularly encourage social participation and interaction by facilitating communication and collaboration among students and with teachers, peer assessment and learning.

Despite the ubiquity and high consumption of such media, surprisingly little research has been done to explore why and how they are used to make meaning and this offers considerable opportunity for new studies (Gleason, 2016). The use of such technologies in formal settings such as educational institutions also offers rich opportunities to study changes to the historical 'imbalance of power within most educational uses of technology' (Selwyn, 2010, p. 71). Twitter and Facebook, like all digital technologies, have particular histories, constraints and affordances that bear on the ways people use them to communicate. Some, like Twitter, have potential to blend private and public spaces and to fuse authorship and readership and create what some see as new expressions and forms of literacy that also offer research opportunities (Greenhow and Gleason, 2012; Stevens *et al*., 2015).

Social media offer ways of exploring the space between institutionally managed systems and non-institutional personal usage; such *liminal* (boundary) areas are seen as fascinating 'third spaces' (Bhabha, 2004; Turkle, 2007; Aaen and Dalsgaard, 2016). However, participants may prefer to keep their social and academic presences separate, so it may be better to explore informal learning using media such as Facebook or Moodle and use other methods for traditional, more formal activity. The division between first (formal) learning spaces and second (informal) spaces can be conscious and functional rather than unconscious, accidental or disruptive; a number of studies have found that students and their teachers often dislike blending study and their social life (Manca and Ranieri, 2013). Research may therefore encounter resistance to using social network technologies for educational uses, as participants may prefer more traditional approaches to learning; in the interests of promoting high-quality research, any such preferences should be incorporated and not overlooked. Additionally, while many Facebook groups are 'open' (public), many are 'closed', and participation may require an invitation (consent) from group members.

23.5 Netography, virtual worlds and social media network software

Developed by Kozinets (2002, 2010), netography (*net*work ethn*ography*) relies heavily on observation and, like traditional ethnography, is an immersive and interpretive exploration of a particular space – in this case of embedded technology. The traditional prolonged immersion of ethnographic research is equally effective in digitally mediated settings, but engagement may be more about following connections than continuous physical presence in one 'space'. As a general rule, researchers should try to engage as fully as possible with the digital environment being studied, but this does not mean that they must participate in every activity in order to conduct meaningful observation. Sometimes full participation is not possible or desirable and total immersion is not essential in ethnography. Nonetheless, such factors may significantly affect how other participants perceive and engage with the researcher, and sufficient expertise with the environment will be a prerequisite for conducting any study, and the persona being portrayed may need to be consistent and coherent across all digital platforms employed.

Netography often uses multifaceted engagement with a setting for studying online hyperlinked resources such as websites, Facebook pages or blogs which have been created by individuals to project their 'fame' (compare with Malinowski, 1922; Munn, 1986). Such studies may include a number of sites and different forms of data collection such as: formal interviews; observing individuals both online and in the physical world (e.g. in cybercafes or game conventions, or whilst physically present but online); watching and chatting to members of the online community being studied, or during casual encounters online and offline. Big data may also be important in this field and Chapter 8 contains an exploration of the implications of its use.

A sense of presence ('being there') in these spaces is likely to be highly individual and conditional upon the user's level of control over the environment (Sheridan, 1992; Ijsselsteijn *et al*., 2000; Sadowski and Stanney, 2002). An *immersive* experience results when presence in an environment is augmented by its apparent overall fidelity to physical reality (Slater and Steed, 2007) and appears when being there is augmented by a total response of 'making sense there' (Schuemie *et al*., 2001; Riva *et al*., 2003). Presence in an online community is often a prerequisite for engagement; new participants often have to be invited or accepted and there may be official gatekeepers or a formal arrangement for gaining access – a rite of passage that existing community members will have gone through so as to be able to participate comfortably in the digital environment. Sometimes it may instead just be a case of building an identity and presence in whatever ways the platform allows, and creating an array of connections over time.

Facebook involves signing up for an account, deciding what photographs and information to upload, which individuals to 'friend' (connect to) and how often to update information; to use Twitter the researcher must decide on a name and what personal information to reveal in the profile and the nature of any 'tweets' or 'retweets' (messages) to be made. These processes are relatively straightforward; but platforms such as World of Warcraft and Final Fantasy are likely to require considerable familiarity and skill with navigation, raids, questing and other features of the environment to facilitate extended contact with other players and acquiring these might take significant time (Sveinsdottir, 2008).

The purpose of both physical world and digital ethnography is not just empirical description but also about developing a theoretically rich description that relates to the particular issue or area of study and connects with wider discussions (Hine, 2015). Many important norms, beliefs, attitudes and behaviours that could be of interest to the researcher can be so taken for granted by participants in a given setting (so 'normal') that they are not remarked or reflected upon by them (Malinowski, 1922). Not all data are therefore accessible through interviews or other traditional means of data recording (indispensable though these are) and so participant observation plays an especially important role. Participant observation within virtual worlds and social network platforms can offer significant opportunities, as the extended timeframe allows the researcher to reflect, revise classifications, assumptions and analysis and to discuss with participants any emerging themes and interpretations. These promote reflection and the revision of assumptions, classifications and analysis. The 'participant' role of the researcher entails searching for unremarked-upon things proactively and being aware that what is observed, even if only a partially glimpsed pattern or behaviour, may be valuable and should be recorded and followed up, even if (perhaps *especially* if) the rest of the community appears to pay it little attention. These processes make short studies more difficult as they require substantial and sustained time and commitment from the researcher, so when planning a project it is important to allow for regular engagement with the community of interest over extended periods (e.g. weeks and months).

Before 1910, there was surprisingly little in the literature of the social sciences about the now commonplace notion of 'community' and the first clear definition of it focused on defining rural communities around villages (Galpin, 1915); in the forty years following, over 100 other definitions appeared (Hillery, 1955). The notion of 'community' in the social sciences grew to include the idea of community as a *value* (see Frazer, 2002) and increasingly embraced overlapping descriptions or ideas that mingled together and became difficult to separate, as, for example, when applied to the study of ideas of solidarity, trust, mutuality, fellowship or conflict. The more recent additions of communities that are 'imagined' or 'virtual' have extended this range and complexity even further (e.g. Anderson, 1983; Rheingold, 2000).

The first task of a researcher studying a community is therefore to clarify what is to be meant in their study by 'community', as communities within digital environments are often only very loosely bounded and dynamic by nature; hence it is important to be able to distinguish who and what is 'in' and 'out' of the study and how this is to be decided. This is what ethnographers would usually think of as identifying the 'field-site', which may traditionally have been one or more geographical areas with relatively clear boundaries (e.g. a village), perhaps with more fluid boundaries for particular social groups. Group 'types' are sometimes also

used this way in researching the digital world and might include professional organizations (communities of practice), or communities that are workgroups, families, friendship groups or diasporic groups or guilds of people who join together for some other activity (e.g. in online games). Researchers may also study online groups, not so much by community membership but by activity, so they may focus on novice participants, or those with longer engagement who may act as managers or leaders, or those who are present but do not otherwise engage ('lurkers'). So a fieldsite can be thought of as a collection of places, individuals or practices that might be physical or virtual or some combination of these; and this kind of study is sometimes called 'connective ethnography' where researchers seek to describe the use of these related sites and explore the connections between them (Leander and McKim, 2003; Hine, 2007; Taylor, 2009).

In netographic research it can be difficult to separate the offline and online lives of individuals, as their online engagement may be deeply enmeshed with their life in the physical world. The researcher cannot then follow the traditional ethnographic methodology of visiting a physical organization or place because online and virtual communities exist in placeless spaces. The online/offline distinction also seems increasingly artificial, as many new digital technologies claim to have a 'social' element. The use of social media is participatory and collaborative by nature, but it is also often complex and intermingled with issues of identity and social relationships (Selwyn and Stirling, 2016). Some suggest that these new media challenge and destabilize the very concept of self (Baudrillard, 2012). One consequence of this is that maintaining boundaries between the research and other aspects of the researcher's life may become difficult; they may have a digital presence that reaches far beyond their project that participants may access. These complexities arise largely because the rise of social network software has shifted Internet use from passive consumption towards active participation, from 'pull' to 'push' behaviours. Conducting participant observation in such settings may be difficult to characterize, as when deciding what 'participation' means and what exactly is being 'observed'; the researcher may be observing interactions which they have in part created, so are in some senses observing themselves (Law, 2004).

However, ethnography never provides a neutral or objective account of what is studied, as the researcher always plays some part in constructing the object of their study, although it is important for them to maintain an awareness of this phenomenon and its likely consequences (see Clifford and Marcus, 1986). Seeking to minimize this effect by adopting only passive observation strategies may prove unworkable, as online communities may detect and censure those online who are not participating (lurking) and using covert observation raises further ethical issues. Visibility is therefore important but does not need to be in every communication medium or to be constant, and it is possible and often acceptable to other participants to observe an online discussion group without posting messages. However, avoiding lurking by posting repeatedly without contributing to the aims and goals of the group would often breach group etiquette and is best avoided.

23.6 Opportunities for research with virtual worlds, social network software and netography

Studies of online communities may involve: study but not participation, which some may regard as somewhat contradictory and ethically questionable; or study with some participation; or study plus offline/online interviews; or study which also includes offline research methods. Deciding on data collection and other appropriate methodology for exploring virtual worlds or digital social networks in educational research, therefore, often involves considering a range of offline and online tools (Fielding *et al*., 2008; Markham and Baym, 2008). Existing community members may have several means of communicating, such as email, face-to-face meetings or through blogging, so a study may not be confined to a single 'site' or software but could include and draw upon a range of online and offline settings. As a result, mixed methods/mixed worlds/mixed media approaches are increasingly common (Johnson *et al*., 2007; Martin *et al*., 2010), although these can exacerbate some research problems, especially of maintaining user engagement (see the case study in Livingstone and Bloomfield, 2010).

Virtual world and social network platforms can provide environments that reflect and include features of 'real life' and are valuable for studying interactions between individuals, especially when we wish to explore contexts where sensitive issues may be the focus. They can be useful for the study of interaction, especially in dynamic, fluid, uncertain or contested contexts, for exploring complex behaviour and for monitoring developments over time. By their nature, virtual worlds, especially the perceptually realistic ones that are increasingly common, offer the researcher an opportunity to exploit a sense of immersion in the created world (a sense of being *there*) and also the sense of a shared experience with others (a sense of *being* there). Both immersion and co-presence have

been recognized as important facilitators of user engagement in a time when media consumers demand more and deeper experiences (Turkle, 2000; Riva *et al*., 2003; Boellstorff, 2008).

Behaviours and attitudes towards others can be explored in such contexts to study contested opinions or beliefs, or individual and self-perceptions in relation to others and how these might change over time. This makes virtual worlds useful places for researching the development of understanding, of perception, of processes where negotiated meaning is important, and of the dynamics that generate consensus and discord. Research should always be focused and highly contextualized but also relate to wider issues of interest; virtual worlds and other social network software are well placed to provide insights into embodiment, selfhood, globalization or learning and many other topics that may be important for those with little interest in digital environments.

Another advantage of virtual worlds lies in exploring contexts where the experimenter neither desires nor is able to exert control over every aspect of the situation and where behaviour is shaped by the agency of participants. Virtual worlds are also highly suited to mixed methods approaches, the development of research methodology and activity design and the exploitation of game theory (Broadribb *et al*., 2009). Technologies such as Second Life, Club Penguin, Facebook or Twitter are collaborative environments suited to 'inclusive' research practices where researchers and subjects engage on equal terms and therefore offer opportunities to develop scenarios of 'ideal speech' (Rybas and Gajjala, 2007; Sheehy, 2010), discussed earlier.

Researchers have explored the influence of social media on campus life, identity, sexuality, relationships and attitudes towards 'others'; some feel that such technologies relax compliance with the social norms that affect inter-person spaces, and that this allows observation of communication with fewer constraints from the social conventions that may appear in traditional face-to-face interactions such as focus groups or interviews (Pitcher, 2016). Social media also have significant potential to disrupt established modes of interaction or hierarchies of authority and power within institutions, to challenge and sometimes remake them, from settings of political activism to those between students and teachers (Selwyn and Stirling, 2016).

However, the open, social and collaborative nature of participatory networks are often seen as also posing challenges for education, implying possible shifts in the roles of learner and teacher, making institutions more porous and raising concerns about untangling collaborative from individual learning and the challenges this presents for fair assessment (Manca and Ranieri, 2016). The effects of social media use in education tend to be quite diverse and sometimes negative; so whilst their use can be correlated with increased student involvement (Junco, 2012) and time on Facebook has been associated with academic success (Labus *et al*., 2015), it has also been found to be inversely correlated with academic progression (Paul *et al*., 2012). Results from different studies sometimes conflict and so can be difficult to integrate, perhaps because different features of particular social media can support diverse forms of involvement and different activities may produce different effects (Matzat and Vrieling, 2016).

Social media can also be used to study a range of demographic and other variables in relation to topics such as teacher networking, student expectations, peer feedback, identity presentation and development, support, maintaining relationships and different phases of transition, in addition to their possible academic uses, which have received relatively little attention. Social media can also present some conundrums for educational institutions; the lack of accountability and other effects of anonymity can lead to cyber-bullying, racial hostility or the promotion of damaging lifestyles. This *disinhibition effect* can occur when anonymity encourages feelings of a 'safe barrier' between perpetrator and victim. Control of content and interaction can also be a 'sharing' that looks like intimacy but is actually a kind of distancing (Burbules, 2016). However, more positive outcomes for classroom engagement and study can also arise from online disinhibition: for example, students may more freely share their academic work, or feelings and problems that they might not reveal offline. Both positive and negative scenarios highlight the need for great care with ethics, where informed consent, data security and participant anonymity may present particular challenges (Rowan-Kenyon *et al*., 2016; see also Chapter 8 of the present volume).

Exposure to conflicting values is likely to be an unavoidable feature of many social media and research into or with them is therefore likely to encounter tensions. In education, this raises questions for the kinds of educational spaces being fostered, as short and perhaps superficial critique from anonymous others may expose students to points of view or ideas that may challenge and disturb them. The 'messy democracy' of social media affords particular kinds of robust interaction for which the researcher and participants need to be prepared; they can be productive and creative but also can be hypercritical and forceful (Burbules, 2016).

Some commentators suggest that: the benefits of social media have been exaggerated and that using such technology actually minimizes opportunities for collab-

oration, as people may work on their own on separate parts of a project; such media therefore lead to greater misunderstanding, less knowledge sharing and less creative or higher-order cognitive processes; using social media such as Facebook is time-wasting and that both students and academics dislike the blurring of their social and professional identities and raise concerns about how they become represented (Manca and Ranieri, 2013; Salmon *et al*., 2015; Stirling, 2016). Twitter is argued by some to have a zero or negative effect on learning (Kucuk and Sahin, 2013; Arabacioglu and Ajar-Vural, 2014), whilst others find its influence positive (Evans, 2014; Ricoy and Feliz, 2016). Kirschner (2015) challenges the idea that Facebook can be useful for formal learning outcomes, citing its inadequacy for academic discussion, knowledge construction and argument; and others find that it encourages narcissism, tribalism and superficiality by connecting individuals with similar views and thus discouraging openness to differing views and the reflective and objective analysis of evidence and extended argument (Manca and Ranieri, 2016). Facebook has also been found to reinforce culturally embedded relations of power distance, as more successful or able students, or students wishing to manifest gratitude or respect, tend to benefit from exchanges with teachers whilst others do not (Tananuraksakul, 2014; Manca and Ranieri, 2016).

Many studies using self-report methods focus on positive implications for learner attitudes, engagement or attendance, whereas studies finding benefits for academic knowledge, understanding or attainment are less common, and those finding negative effects tend to have used objective data (Tess, 2013). Opportunities exist for more empirical studies of direct benefits for 'hard' outcomes such as academic progression, knowledge, understanding or attainment when comparing settings with and without the use of social media integration (Ricoy and Feliz, 2016). Other potential areas of study include implicit and explicit institutional policies and traditional pedagogic and role expectations, which may be especially interesting in cultural settings where the maintenance of social harmony is important.

23.7 Ethics

How should we treat information posted online, such as images from a mobile phone or information perhaps of a personal nature intended for family and friends but not for others? This could be important when individuals may not have a strong understanding of the possible long-term ramifications of such posting, even though the data are potentially discoverable and therefore already 'public'. It is no excuse for the researcher to say that participants should know the terms of service and functionality of platforms such as Twitter or Facebook when they open an account. It is likely that many experienced users of Facebook remain unaware that people who have not 'friended' them can nonetheless access their photo albums, or that images uploaded to social media often include embedded metadata with details including author, location, time and much more (see Raynes-Goldie, 2010).

Social media and virtual worlds present particular problems for traditional frames of reference for ethical research, especially when children or young people are involved. It may be unclear who the participants actually are, to whom the data belongs, what data can legitimately be regarded as 'public' or 'private'; and the possible consequences for participants and their networks now and in the future may be unknowable. The main ethical issues here concern informed consent, the vulnerability and individual risk with online identities, the public/private nature of communication, security and confidentiality. There is no single template for ethical research and Internet-mediated research is no exception; the context and nature of the study will influence ethical considerations. For example, covert observation is often frowned upon because it violates the principle of informed consent but it may be ethically acceptable and individuals' informed consent may not be necessary in online or other contexts where data exist within the public domain and where the risk of harm to users is low (Steven *et al*., 2015). Other covert or 'unobtrusive' methods may rely on publicly available data, as one kind of unstructured observation of things that the researcher might not be able to obtain or ask about directly. Such methods may be helpful and ethical where participants could find it difficult to give authentic or honest answers, perhaps because these may be socially undesirable, or sensitive, or where not employing unobtrusive approaches would encourage only 'diplomatic' responses.

Unobtrusive methods may not engage participants, or solicit comments from them and therefore use the role of lurker or voyeur. Users may not have agreed to participate (and may not have been asked) and so anonymizing the data becomes especially important. Examples of unobtrusive ways of collecting social media data may include: online discussion forums; Google Trends data (see www.google.com/trends); Facebook and Twitter postings; YouTube videos; and downloaded 'chats' from message rooms, community sites or user forums. However, such 'found' or 'non-reactive' data may easily become 'reactive' if reinserted into the digital setting, such as, for example, when presented to participants to ask them what they made of it.

It may be ethically problematic to analyse found data in the absence of participant consent, unless interpretations are depersonalized and carefully justified. However, depersonalizing is no longer a strong guarantee of anonymity, as '[t]oday, the private man is a public entity, even a public display, that he controls only partly' (International Council on Human Rights Policy, 2011) because

> in an international and pervasive network (e.g. the Internet) that is persistent in its records, and increasingly searchable across indexes, databases and other taxonomies, ultimately every interaction online has the potential to be traceable, either now or in the future.
>
> (Henderson *et al*., 2013, p. 551)

One solution might be to adopt the principle of 'non-alienation' (Bakardjieva and Feenberg, 2000), where data may not be removed from someone's control or used for things they were not aware of without their explicit permission, which employs the principle of consent being ongoing, where participants are consulted at all stages of the research so as to have more control over data collection, analysis and reporting of research (Henderson *et al*., 2013; Ramírez and Palu-ay, 2015).

The notion of informed consent is therefore potentially complex with all social digital media, even when the risk to participants may be outweighed by the potential benefits to knowledge about the field of study (Pitcher, 2016). Nonetheless covert or unobtrusive data collection is common (arguably necessary) in some disciplines even when it 'potentially poses a substantial threat to those who are involved or have been involved in it' (Lee, 1993, p. 4), or when those studied may view the research as somehow undesirable (Van Meter, 2000). This is because the research may be deemed to have overriding benefits for the good of wider society (e.g. when studying illegal activity). Chapter 8 of the present volume discusses further a range of ethical issues in Internet research.

23.8 Guidelines for practice

A range of issues have to be considered when conducting research using virtual worlds and social networks (Moschini, 2010; Hine, 2015) and each may be more or less prominent depending on the setting and nature of the research:

- The holistic approach of netography means remaining alert for unanticipated acts of meaning-making and for how activities make sense to the individuals engaged in them (Lewis and Allan, 2005).
- Fieldsites are rarely online or offline but are fluid, and researchers may need to engage with both; do not assume the existence of boundaries (Johnson *et al*., 2007; Martin *et al*., 2010).
- Explore all stages and forms of engagement and use a range of tools for recording and interpretation (Broadribb *et al*., 2009).
- Virtual worlds and social media sites have customs and norms that mediate communication, but to their users may be simply 'the way things are'; adopting the perspective of the stranger can help to understand why this may be so or could be otherwise (Malinowski, 1922).
- Expect a variety of different media and experiences across multiple platforms.
- Expect uncertainty: digital spaces often do not provide a single, verifiable or 'objective' reality.
- Allow sufficient time: familiarity and expertise may be needed to engage with the community and gain acceptance.
- Ensure anonymity of data, whether participants request it or not; no one can know what may happen in future, and identifying an individual may also identify their social networks once data and analysis are in the public sphere (Henderson *et al*., 2013).
- When participating within a particular group, avoid taking sides; avoid conflict.
- Remain alert to the way the researcher's own agency intrudes in the process of creating an authentic, rich and thick account of the 'messy reality' of the digital cultural space as perceived and understood by participants (Law, 2004).
- Be aware that study participants may have access to the researcher's own online identity; consider whether and how this may influence the study and plan accordingly.
- Ideally all forms of misrepresentation or deception are best avoided; they are antithetical to openness and trust. Covert observation is a form of deception that cannot be guaranteed to be free of potential harm because we do not know what may be observed. However, some kinds of research may be less reliable or even impossible without some form of deception, and the balance and degree of any potential harm/benefit requires careful consideration and justification throughout (Steven *et al*., 2015).

In addition to the usual decisions to be made when designing a research project, using a virtual world or social network will require it to be set up and managed and may need someone with technical expertise to help decide whether to make use of an existing commercial product or to create a purpose-built environment. The

former is more straightforward but the options available for customization may be more limited, so it is important to be clear in advance which features are required for the study, what data need to be collected, and how this will be done. A customized platform may be the ideal, particularly in the case of a virtual world, but the resource implications may be significant and, in the interests of viability, it may instead be worth adjusting the research methodology to allow the use of a commercial product.

A prerequisite for research in any online environment is sufficiently high-quality online availability, and a prudent researcher will check this and software/system requirements at an early stage. The researcher must also fully understand the protocols, etiquette and common practices of the fieldsite(s) sufficiently for them to engage and blend in with the community they wish to study. This may require technical and personal preparation before beginning the study, and the time and energy that will need to be dedicated to such 'acculturation' should not be underestimated.

As with all contexts for research, enquiry into digital artefacts such as virtual worlds, digital social networks, online forums, blogs or wikis should be driven primarily by research questions, not by decisions about which methodology to adopt, and no particular mode of study should be automatically privileged above others. Instead, what matters is a clear and careful link between the research question and the methodological design of the research. This is often an iterative process and an effective researcher will always be open to the possibility of adjusting (or even sometimes discarding) his/her original research question as exploration proceeds and new insights appear. Exploration and insights offer valuable opportunities to redirect the research to exploit them, and research should always respond to the pressure of evidence (Malinoswki, 1922).

By their nature, virtual worlds and social networks lend themselves to projects which do not require participants or researchers to be physically located near each other. However, much existing research with these technologies focuses on single groups of educators or students or members of a specific university, so a project with more breadth than this would have increased value. Collaborative affordances may then become important if using a research team, and researcher collaboration can be facilitated by technologies such as Google docs or Dropbox to share documents, spreadsheets and databases, and work on them to develop understandings or dynamic trend analysis using notes, 'mind-maps', graphs, charts or tools such as Gapminder.

Challenges can emerge when creating common protocols for a research team, especially if individuals are based in different institutions or countries, as these may apply different procurement constraints, use dissimilar and sometimes incompatible IT infrastructures, or have policies with very different embedded institutional and cultural assumptions about the nature of academic roles and responsibilities (e.g. for academics and instructors or for students and research participants). When using virtual worlds, having computers with appropriate specifications will also be important (check the software's website), along with sufficient bandwidth and safe passage through institutional network firewalls, so discussions with institutional IT managers are important when planning a study.

Whilst participant enthusiasm is common, it should not be assumed and some studies have found to their cost that participants who are initially enthusiastic find that time constraints, the relative complexity of virtual worlds and bandwidth demands (which can create operational slowness) can prove a disincentive for continuing engagement (see Jarmon *et al*., 2009). Despite the visual and conceptual allure of virtual worlds and the popularity of social media, successful participant recruitment may require careful preparation (Fetscherin and Lattemann, 2007). Consider who might facilitate recruitment beyond known contacts, to include local authorities, professional organizations, universities or employers. It may be useful to set up a website to explain the research and provide information and documentation, and this will also facilitate engaging with, and recruiting, potential participants.

23.9 Data

Because virtual world and social media fieldsites may be large, diffuse and engage with different media, it can be hard to establish the target population or get a sense of how representative a given sample may be (Hine, 2015). In many virtual environments, events and objects may also be connected to other online technologies such as blogs, wikis, questionnaires, rating systems, databases, etc. and increasingly to in-world tools. In such circumstances, careful planning of data-collection strategies is essential.

Communication in virtual worlds via text, chat, voice and signing can all be captured for later analysis via recording and transcription and, because data can be time-stamped, it is possible to compare the outcomes from analysis from these different communication channels for data triangulation (Martin and Vallance, 2008). By providing a scaffolded vocabulary, researchers may also more easily collect data from individuals with communication difficulties, where the slower pace of interaction and reduced amount of data involved in communication in these environments (such as lack of subtle

facial or body-language communication) may be advantageous (Ravenscroft and McAllister, 2006). These data can be converted to numbers for quantitative analysis or can be used to develop a richer understanding through interpretative, phenomenological analysis, which is a useful form of qualitative analysis when we are interested in describing how people negotiate, understand and make sense of the world. Both quantitative and qualitative data need to be used in light of the perspective adopted, the kind of information collected and the assumptions and objectives of the research.

Field notes are an important element of data collection in netography but they should be more than just descriptive accounts; they can be early interpretations of what happens, notes about feelings, ideas about what these might mean, why things happened or did not happen, 'to do' lists, as well as notes of frustrations or puzzles. They might include sketches to suggest meaning, proximity or relationships, or be loosely organized reflections and ideas for later use and reflection. The researcher may wish also to blog or tweet about the research or post things on Facebook or Instagram to encourage further interaction and debate, although the consequences may be some potentially awkward decisions about what to reveal and concerns about 'over-sharing' that may preempt later and more thoughtful analysis (Boellstorff, 2008; Boellstorff *et al*., 2012; Hine, 2015).

Screen shots, field notes and other data should be numbered, dated, time-stamped and written up immediately after an intervention. Making notes with a word processor can make it easier to search for keywords later, and audio or video recording of sessions can be useful for future analysis and allow the researcher to concentrate on making observations during sessions. Virtual worlds and social media allow researchers to collect more participant data than are otherwise possible, by capturing text, chat and user presentations from the in-world environment and by recording activity. It is important to establish whether the chosen online setting will automatically capture data from features like chatlogs and, if so, this should be tested to make sure the date and time are always recorded.

It may not be immediately clear what valuable information is within such data, as it will often contain intermingled issues, but it cannot be collected retrospectively, so capturing it as work progresses is important. Each specific event should be recorded with a unique filename so that data can be correctly sequenced and matched together later, as it might be extremely difficult to do this retrospectively. Collecting data from online environments is relatively easy and can quickly generate large amounts of material, so all relevant information needs to be kept together as it is collected. Leaving this task until later may be unwise, as it may then be difficult to remember which data goes where and to what other data it is related.

Before choosing software to be used for recording data it is important to consider:

- whether it will meet the research needs;
- whether it can be supported by and run effectively on the computer system to be used (e.g. PC, Mac, smartphone);
- whether enough storage is available, and how long recordings will usually be and how many of these there might be. How much detail is important: do video recordings need to be full-screen or high-definition? These considerations will affect the size of recorded files, so using a large-capacity external drive may be necessary;
- whether advanced editing features will be needed or just a basic package;
- testing all software thoroughly with the system on which it will be used before adopting or purchasing it, and running extensive tests before the research begins, to minimize the risk of technical problems.

Visual representations such as screenshots can be useful when combined with other data, either to illustrate an observation in the field notes or to remind the researcher of some important event or pattern of behaviour. Using visual records from online environments, whether single images or video, requires careful ethical consideration and participants need to be aware that such records will be collected and to have given prior and informed consent. Individuals must also not be identifiable from visual data; care must be taken to anonymize any that are used by removing identifying elements (see also Chapter 31 of the present volume).

Visual data are not always available, as some online environments can be entirely composed of text, or have only text chat as their means of communication. For many others, voice chat is increasingly an inbuilt feature, but where not it is common for online communities to employ additional software, effectively a phone service over the Internet known as Voice over Internet Protocol (VoIP). The researcher may therefore need to make enquiries within the community to be studied to determine whether VoIP is used, and then adopt it to augment the research data by capturing important communications, ambient sounds or music.

Participants in an online environment may also make use of associated separate communication media, but the individuals in any of these other communications may not necessarily be frequent, or the same users as found in the main study site, and careful notes will help with

mapping these different audiences and sub-groups to ensure clarity and avoid unhelpful confusion during analysis. As is the case for all data, keeping notes of filenames and associated other data will be essential, so as to know which participants were present, the time, the event(s) and location and any other relevant details. Finally, offline gatherings such as refreshment places, conventions and workshops are sometimes associated with, and used by, particular online communities and can be valuable sites for gathering additional data and meeting known or new members.

23.10 Conclusion

This chapter has argued that a range of emergent, contentious or sensitive topics can be usefully explored using virtual environments and networked social media and that these present both opportunities and challenges. Virtual worlds and increasingly other social media can offer high levels of interaction and embodiment and heightened immersion via technologies such as Oculus Rift, PlayStation VR or HTC Vive. This combination of technologies offers rich communication affordances and some unique opportunities for data capture,[2] including the ability to record sound, chat and text as well as contemporaneous images and video in real time.

In such research settings there may be a need to scrutinize what is understood by ethnographic or qualitative approaches such as participant observation, focus groups and interviewing. The traditional assumptions that are made when deploying methodological tools may need to be revisited to verify whether they are still applicable in these digitally blended communities, in which alternative values and different articulations of reality are to be found, where participants may alter their displayed embodiment at will, and in which we may simultaneously conduct our research with individuals in different locations around the world. By exposing the researcher and practitioner to new constructions, expressions and transformations of identity, reality and community, virtual worlds and other social network software offer many unique opportunities to re-examine the nature of community and self across the virtual/physical world and to rethink, refine and improve existing pedagogy and research methodology and instrumentation.

The companion website to the book provides worked examples of virtual world, social media and netography research, together with website references for the topics addressed in this chapter, virtual world research methods and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: www.routledge.com/cw/cohen.

Notes

1 The term ‘virtual world’, as used here, presumes the presence of human users, although strictly speaking even when empty of participants such an environment may remain a virtual world. Part of the reason for the ongoing debate about what is and what is not a virtual world, and whether these things should be given a different name, is that the use of virtual world features is becoming more common in the technologies of the World Wide Web, in which one increasingly sees visually realistic three-dimensional virtual environments on websites, discussion forums, blogs, chat rooms and social network sites where user involvement is mediated by avatars.

2 Many commercial companies invest a great deal of time and money acquiring and analysing data from social media. These companies often use expensive specialist companies to do this work but are sometimes prepared to offer substantially lower prices to students or academic institutions, so check to see if your institution has or can secure this. There are also some relatively straightforward methods of collecting social media data that can be used by researchers for free or at minimal cost.

Twitter: Entrepreneur (www.entrepreneur.com/article/242830); The Chorus Project (http://chorusanalytics.co.uk) (see also Tweetcatcher).

Facebook: Graph (www.facebook.com/graphsearcher); Statista (www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide) (check to see if your institution provides access).

VennMaker (a mapping tool for collecting and analysing data in social network analysis) (www.vennmaker.com/?lang=en).

Some straightforward guidance is also available on obtaining data from most social media software at WikiHow (www.wikihow.com/Main-Page).

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Part 4
Methods of data collection

This section moves to a closer-grained analysis of instruments for collecting data, how they can be used, how they can be constructed, what are their strengths and weaknesses, how to work with them and what considerations have to be borne in mind when deciding on the most appropriate choice of instrumentation. We identify eight main kinds of data-collection instruments, with many variants included in each: questionnaires (with greater coverage of online questionnaires); interviews; observations; tests; personal constructs (written by Richard Bell); role-playing (written by Carmel O'Sullivan); an entirely new chapter on using secondary data in educational research; and an updated chapter on visual media in educational research. We have expanded our discussion of material from the previous editions, particularly in respect of questionnaire design, online interviews and the increasing use of visual media in educational research, including photo-elicitation and video research and the ethical issues surrounding these.

Selecting the instrument(s) for data collection, like deciding on methodologies in Part 3, is not a matter of preference, arbitrary or automatic decision making, but, like other aspects of research, is a deliberative process in which the key is the application of the notion of *fitness for purpose*. The intention of this part is to enable researchers to decide on the most appropriate instruments for data collection, and to carry out the practical, careful design and use of such instruments. The strengths and weaknesses of these instruments are set out, so that decisions on their suitability and the criterion of *fitness for purpose* can be addressed. Hence this part not only introduces underlying principles that underpin instruments, but also offers sound, tested, practical advice for their usage, to enable researchers to gather useful and usable data. There is greater coverage of conducting research that involves children. We provide practical advice to researchers who are considering these instruments for data collection and how to use them, what safeguards to address, what challenges they might face and how to overcome them.

The companion website to the book provides supporting materials and PowerPoint slides for Part 4. This resource can be found online at: www.routledge.com/cw/cohen.

Questionnaires

CHAPTER 24

The chapter follows a sequence in designing a questionnaire, thus:

- ethical issues
- planning the questionnaire
- types of questions
- avoiding pitfalls in question writing
- sequencing questions and the layout of the questionnaire
- covering letters/sheets and follow-up letters
- piloting the questionnaire
- administering questionnaires
- processing questionnaire data

24.1 Introduction

Questionnaires offer benefits of standardized and open responses to a range of topics from a large sample or population. They can be cheap, reliable, valid, quick and easy to complete. The field of questionnaire design is vast. This chapter provides a straightforward introduction to its key elements, indicating main issues to be addressed, some important challenges and how they can be resolved. We advise readers to take this chapter together with the chapters in this book on surveys, sampling and interviewing (Chapters 17, 12 and 25 respectively). Chapter 18 (Internet surveys) addresses important material on online questionnaires, and we advise readers to consult that in detail, as we do not address such questionnaires in the present chapter, other than in passing. Part 5 concerns data analysis and this can include analysis of quantitative and qualitative data from questionnaires.

We suggest that researchers may find it useful to work through these issues in sequence, though, clearly, a degree of recursion is inevitable. Whilst many questionnaires are devised electronically, with templates and attractive layouts, these do not obviate the need for careful consideration of the issues addressed in this chapter, as, regardless of the software available, the researcher has to take a wide variety of decisions on all aspects of the questionnaire.

The questionnaire is a widely used and useful instrument for collecting survey information, providing structured, often numerical data, able to be administered without the presence of the researcher and often comparatively straightforward to analyse. These attractions have to be counterbalanced by the time taken to develop, pilot and refine the questionnaire, by the possible unsophistication and limited and superficial scope of the data that are collected and by the likely limited flexibility of response (though, as Wilson and McLean (1994, p. 3) observe, this can frequently be an attraction). The researcher will have to judge the appropriateness of using a questionnaire for data collection, and, if so, what kind of questionnaire it should be.

24.2 Ethical issues

The questionnaire will always be an intrusion into the life of the respondent, be it in terms of time taken to complete the instrument, the level of threat or sensitivity of the questions, or the possible invasion of privacy. Questionnaire respondents are not passive data providers for researchers; they are subjects not objects of research. There are several ethical sequiturs that flow from this; we introduce these below, and advise readers also to review Chapters 7 and 8 of the present volume.

Respondents cannot be coerced into completing a questionnaire. They might be strongly encouraged, but the decision whether to become involved and when (and if) to withdraw from the research is entirely theirs. Their involvement in the research is likely to be a function of:

- their *informed consent* (see Chapter 7);
- their *rights to withdraw* at any stage or *not to complete* particular items in the questionnaire;
- the potential of the research to improve their situation (the issue of *beneficence*);
- the guarantees that the research will not harm them (the issue of *non-maleficence*);
- the guarantees of *confidentiality*, *anonymity* and *non-traceability* in the research;
- the degree of *threat* or *sensitivity* of the questions (which may lead to respondents' over-reporting or under-reporting (Sudman and Bradburn, 1982));

- factors in the questionnaire itself, for example, its coverage of issues, its ability to catch what respondents want to say rather than to promote the researcher's agenda, i.e. the avoidance of bias and the assurance of validity and reliability in the questionnaire – *methodological rigour* and *fairness*. Methodological rigour is an ethical not simply a technical matter, and respondents have a right to expect reliability and validity;
- the *reactions* of the respondents. For example, they may react strongly if they consider an item to be offensive, intrusive, misleading, biased, misguided, irritating, inconsiderate, impertinent or abstruse.

These factors impact on every stage of the use of a questionnaire, suggesting that attention has to be given to the questionnaire itself, the approaches made to the respondents, the explanations given to the respondents, the data analysis and the data reporting.

24.3 Planning the questionnaire

The overall plan

At the preliminary stage of design, it can sometimes be helpful to use a flow chart to plan the questionnaire. In this way, researchers are able to anticipate the type and range of responses that their questions are likely to elicit. We set out a staged sequence for planning a questionnaire in Figure 24.1.

Within these stages there are several subcomponents, and this chapter addresses these. Further, though these stages are set out in a sequence, the process is recursive as the questionnaire design and refinement take place. These are introductory issues, and the remainder of this chapter takes each of these and unpacks them in greater detail.

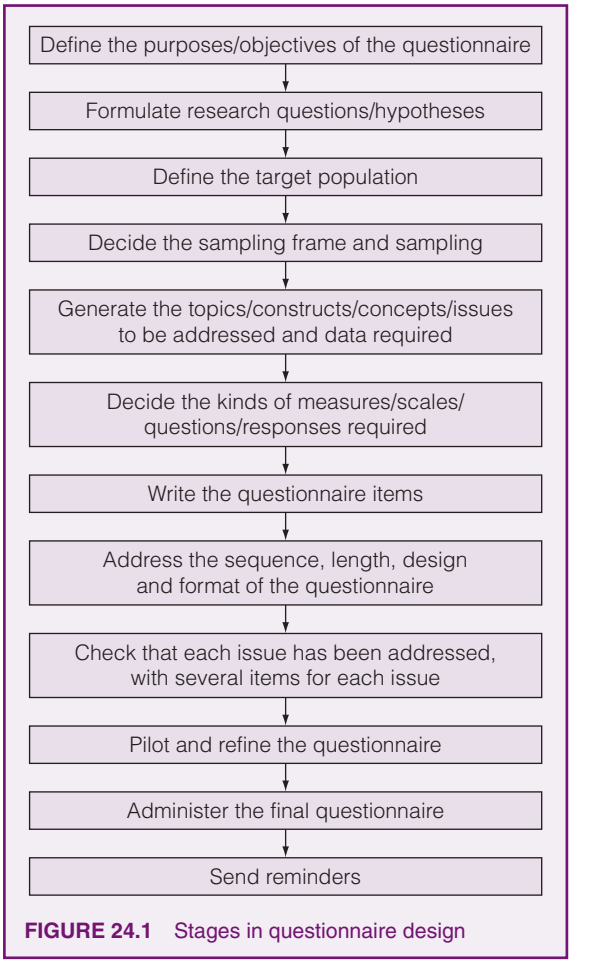


FIGURE 24.1 Stages in questionnaire design

Operationalizing the questionnaire

The process of operationalizing a questionnaire is to take a general purpose or set of purposes and turn these into concrete, researchable fields about which actual data can be gathered. First, a questionnaire's general purposes must be clarified and then translated into a specific, concrete aim or set of aims. Thus, 'to explore teachers' views about in-service work' is somewhat nebulous, whereas 'to obtain a detailed description of primary and secondary teachers' priorities in the provision of in-service education courses' is reasonably specific. Write the purposes of the questionnaire and review them to make them concrete, focused and specific.

Having decided upon and specified the primary objectives of the questionnaire, the second phase of the planning involves the formulation of the research questions to be answered and/or hypotheses to be tested. We refer the reader to Chapter 10 here, noting that the research questions, deriving from the overall purposes, must be concrete, specific and focused, enabling concrete answers to be given.

Then follows the identification of the target population and sampling, as this influences the framing of the questions, their terminology, their level of demand and the medium for administering the questionnaire (e.g. post, face-to-face interview, Internet and email, drop-off).

Then follows the identification and itemizing of subsidiary topics that relate to its central purpose. In our example above, subsidiary issues might well include: the types, contents, location, timing, design and financing of courses.

After the identification and itemization of subsidiary topics there follows the formulation of specific information requirements relating to each of these. For

example, with respect to the type of courses required, detailed information would be needed about the duration of courses (one meeting, several meetings, a week, a month, a term or a year), the status of courses (non-award bearing, award bearing, with a certificate, diploma, degree granted by a university) and the orientation of courses (theoretically oriented involving lectures, readings, etc., and/or practically oriented involving workshops and the production of curriculum materials).

What we have in the example, then, is a move from a generalized area of interest or purpose to a very specific set of features about which direct data can be gathered.

Wilson and McLean (1994, pp. 8–9) suggest an alternative approach, which is to identify the research problem, then to clarify the relevant concepts or constructs, then to identify what kinds of measures (if appropriate) or empirical indicators there are of these, i.e. the kinds of data required to give the researcher relevant evidence about the concepts or constructs, for example, their presence, their intensity, their main features and dimensions, their key elements etc. We have included such matters in Figure 24.1.

What unites these two approaches is their recognition of the need to ensure that the questionnaire: (a) is clear on its purposes; (b) develops concrete research questions which lead into the formulation of the questionnaire items; (c) is clear on what needs to be included or covered in the questionnaire in order to meet the purposes and research questions; (d) is exhaustive in its coverage and inclusion of items; (e) asks the most appropriate *kinds* of question; (f) elicits the most appropriate *kinds* of data to answer the research purposes and sub-questions; asks (g) and for empirical data. We address these points below.

Planning with the data analysis in mind

When planning a questionnaire it is important to plan so that it is set up – structured – in such a way that the data analysis can proceed as planned. For example, if the researcher wishes to conduct multiple regression (e.g. to find out the relative weights of a range of independent variables on a dependent variable), then both the independent and dependent variables must be included in the questionnaire and must be ratio data (discussed below).

For example, let us imagine that the researcher is investigating the relative strengths of reasons (independent variables) why undergraduate students take part-time jobs (dependent variable) (cf. Morrison and Tam, 2005). She asks the respondents about the level of importance of each of the following reasons, awarding a mark out of 10 for each reason, where 0 = 'of no importance' and 10 = 'of very great importance':

- meet necessary study expenses;
- meet living expenses;
- purchase better consumer products;
- support entertainment expenses;
- for extra money to spend;
- support family expenses;
- gain job experience;
- fill in spare time;
- influence of peer group.

She can then conduct a multiple regression to see the relative importance of each of these independent variables on the dependent variable (e.g. see Chapter 42).

If the researcher wishes to conduct factor analysis then the variables must be at the ratio level of data (discussed below). If structural equation modelling is required then both variables and factors have to be calculated, and these must be able to be calculated in the questionnaire. If simple frequencies, percentages and correlations are to be calculated then the questions must be framed in such a way that they can be calculated. This is a statement of the obvious, but, in our experience, too many students neglect the obvious. As Voltaire remarked, 'commonsense is not so common'.

A researcher may not wish to conduct such high-level data analysis, and often simple frequencies will suffice and can be very persuasive. This, too, can suggest causality (though not prove it – see Chapter 6), or at least correlation. Let us imagine that the researcher is looking into the effects of communication on leadership in a secondary school (160 teachers). She asks three simple questions:

1 Generally, how effective is the overall leadership in the school (tick one only):
☐ Good ☐ Not Good

2 Generally, how effective is the principal's communication in the school (tick one only):
☐ Good ☐ Not Good

3 Generally, how willing to communicate is the school principal (tick one only):
☐ Good ☐ Not Good

These simple dichotomous questions require respondents to come to a judgement; they are not permitted to 'sit on the fence', they have to make up their minds. In tabular form, the results could be presented as shown in Table 24.1 (fictitious figures) (cf. Hellevik, 1988).

TABLE 24.1 CROSSTABULATION OF RESPONSES TO TWO KEY FACTORS IN EFFECTIVE LEADERSHIP

Effective leadership	*Principal's communication*	*Willingness to communicate*	*Frequency (% rounded)*
Good	Good	Good	45 (28.1%)
Good	Good	Not good	15 (9.4%)
Good	Not good	Good	10 (6.2%)
Good	Not good	Not good	12 (7.5%)
Not good	Good	Good	3 (1.9%)
Not good	Good	Not good	12 (7.5%)
Not good	Not good	Good	5 (3.1%)
Not good	Not good	Not good	58 (36.3%)
Total			**160 (100%)**

In Table 24.1, using fictitious figures, 'effective leadership' is reported by eighty-two respondents (51.2 per cent) (45+15+10+12); 'not good' leadership is reported by seventy-eight respondents (48.8 per cent) (3+12+5+58). Table 24.1 indicates that for 'good' leadership to be present in its strongest form, the factors 'principal's communication' and 'willingness to communicate' are required to be present and 'good', and that if either or both of these factors is 'not good' then 'good' management drops dramatically.

The point to be made here is that the questionnaire is designed – set up – with the analysis in mind; the researcher knows in advance how she wants to analyse the data, and the structure and contents of the questionnaire follow from this.

Considering the demands on the respondent

It is important to avoid putting too much strain on the respondent, for example, in relying on their recall (Champagne, 2014), in the sensitivity of the issue, in the time taken to complete the questionnaire, in understanding the question. Too much strain can lead to poor-quality or incorrect responses, non-response or quitting. Denscombe (2014) comments that having too long a questionnaire can lead to respondent fatigue. He notes (pp. 172–3) that completing a questionnaire can be mentally demanding, and researchers should consider the burden of effort and demand placed on the respondent: too much and they will not take part, withdraw partway through or give responses which are 'satisficing' (discussed later). Simply because the researcher is keen to acquire data does not mean that the respondent is interested in or concerned about the matter, hence the researcher needs to motivate the respondent and make the topic interesting, meaningful, of concern and motivating. A topic which really concerns the respondent is likely to have a better response than one which is perceived to be irrelevant or of no importance or interest to him/her.

Structured, semi-structured and unstructured questionnaires

Though there are many types of questionnaire, there is a simple rule of thumb: the larger the size of the sample, the more structured, closed and numerical the questionnaire may have to be, and the smaller the size of the sample, the less structured, more open and word-based the questionnaire can be.

The researcher can select several types of questionnaire, from highly structured to unstructured. If a closed and structured questionnaire is used, enabling patterns to be observed and comparisons to be made, then the questionnaire will need to be piloted and refined so that the final version contains as full a range of possible responses as can be reasonably foreseen. Such a questionnaire is heavy on time early in the research; however, once the questionnaire has been 'set up' then the mode of analysis might be comparatively rapid. For example, it may take two or three months to devise a survey questionnaire, pilot it, refine it and set it out in a format that will enable the data to be processed and statistics to be calculated. However, the 'trade-off' from this is that the data analysis can be undertaken fairly rapidly; we already know the response categories, the nature of the data and the statistics to be used; it is a matter of processing the data (e.g. by computer) and analysing and reporting the results.

It is perhaps misleading to describe a questionnaire as being 'unstructured', as the whole devising of a questionnaire requires respondents to adhere to some form of given structure. That said, between a completely open questionnaire that is akin to an open invitation to 'write what one wants' and a completely

closed, completely structured questionnaire, there is the powerful tool of the semi-structured questionnaire. Here a series of questions, statements or items are presented and the respondents are asked to answer, respond to or comment on them as they wish. There is a clear structure, sequence and focus, but the format is open-ended, enabling respondents to reply in their own terms. The semi-structured questionnaire sets the agenda but does not presuppose the nature of the response.

24.4 Types of questionnaire items

There are several kinds of question and response modes in questionnaires, including: dichotomous questions; multiple-choice questions; rating scales; constant sum questions; ratio data; and open-ended questions. These are considered below (see also Wilson, 1996). Questions must be straightforwardly presented, comprehensible at first glance, concrete, specific, unambiguous and able to be answered, which means that assumptions are made that: (a) the respondents know the answers and have an opinion; (b) the demand and effort placed upon them are not too great and that they can actually articulate their response; (c) their recollection and memory are reliable and so on. It is essential that question types are fit for purpose (Champagne, 2014), being suitably focused and concrete (rather, than, for example, being too general and abstract), yielding useable and relevant data, measuring what they are intended to measure and avoiding questions to which the researcher already knows the answer. We consider these and other points below.

Open-ended questions

The open-ended question is an attractive device for smaller-scale research or for those sections of a questionnaire that invite an honest, personal comment from respondents in addition to ticking numbers and boxes. Here the questionnaire puts the open-ended questions and leaves a space (or draws lines) for a free response. Open-ended responses might contain the 'gems' of information that otherwise might not be caught in the questionnaire. Further, it puts the responsibility for, and ownership of, the data much more firmly into respondents' hands.

It is useful for the researcher to provide some support for respondents, so that they know the kind of reply being sought. For example, an open question that includes a prompt could be:

'Please indicate the most important factors that reduce staff participation in decision making';
'Please comment on the strengths and weaknesses of the mathematics course';
'Please indicate areas for improvement in the teaching of foreign languages in the school'.

An open-ended question might frame the answer, just as the stem of a rating scale question might frame the response given. However, an open-ended question can catch the authenticity, richness, depth of response, honesty and candour which, as is argued elsewhere in this book, are hallmarks of valid qualitative data.

Oppenheim (1992, pp. 56–7) suggests that a sentence-completion item is a useful adjunct to an open-ended question, for example:

> Please complete the following sentence in your own words:
>
> An effective teacher …
>
> or
>
> The main things that I find annoying with disruptive students are …

Open-endedness also carries problems of data handling: too many answers to be able to summarize easily; data overload. If one tries to convert opinions into numbers (e.g. so many people indicated such-and-such a degree of satisfaction with the new principal's management plan) – quantitizing qualitative data – then maybe the questionnaire should have used rating scales in the first place. Further, it might well be that the researcher here is in danger of violating one principle of word-based data, which is that they are not validly susceptible to aggregation, i.e. trying to bring to word-based data some principles of numerical data, borrowing from quantitative, positivist methodology to inform a qualitative, interpretive methodology.

Further, if a genuinely open-ended question is being asked, responses may not bear such a degree of similarity to each other to enable them to be aggregated too tightly. Open-ended questions make it difficult for the researcher to make comparisons between respondents, as there may be little in common to compare. Moreover, to complete an open-ended questionnaire takes much longer than placing a tick in a rating scale response box; not only will time be a constraint here, but there is an assumption that respondents will be sufficiently or equally capable of articulating their thoughts and committing them to paper or to the box on a screen.

In practical terms, Redline *et al*. (2002) report that using open-ended questions can lead to respondents

overlooking instructions, as they are occupied with the more demanding task of writing in their own words than reading instructions.

Despite these cautions, an open-ended question is a window of opportunity for the respondent to shed light on an issue, and thus has much to recommend it.

Open-ended questions are useful if the possible answers are unknown or the questionnaire is exploratory (Bailey, 1994, p. 120), or if there are so many possible categories of response that a closed question would contain an extremely long list of options. They also enable respondents to answer as much as they wish, and in their own words, and are particularly suitable for investigating complex issues, to which simple answers cannot be provided. They can generate rich data. Open questions can be useful for generating items that will subsequently become the stuff of closed questions in the final version of a questionnaire (i.e. part of a pre-pilot). Krosnick and Presser (2010, p. 267) note that open items often provide more valid and reliable responses than closed items, but that respondents are more likely to opt for a 'don't know' response rather than take the time to complete an open question.

Open questions enable participants to write a free account in their own terms, to explain and qualify their responses and avoid the limitations of pre-set categories of response. On the other hand, they can lead to irrelevant and redundant information; they may be too open-ended for the respondent to know what *kind* of information is being sought; they may require much more time from the respondent to enter a response (thereby leading to refusal to complete the item); respondents may have difficulty in articulating their thoughts; and open-ended questions may make the questionnaire appear long and discouraging. With regard to analysis, the data are not easily compared across participants, and the responses are difficult to code, classify and analyse.

Closed questions

Closed questions prescribe the range of responses from which the respondent may choose. Highly structured, closed questions are useful in that they can generate frequencies of response amenable to statistical treatment and analysis. They also enable comparisons to be made across groups in the sample (Oppenheim, 1992, p. 115). They are quicker to code and analyse than word-based data (Bailey, 1994, p. 118), and, often, they are directly to the point and deliberately more focused than open-ended questions, helping the respondent to answer easily, as response categories are provided; processing vast quantities of word-based data in a short time frame is extremely demanding.

If a site-specific case study is required, then qualitative, less structured, word-based and open-ended questionnaires may be more appropriate as they can capture the specificity of a particular situation. Where measurement is sought then a quantitative approach is required; where rich and personal data are sought, then a word-based qualitative approach might be more suitable.

In general, closed questions (dichotomous, multiple choice, constant sum and rating scales) are quick to complete and straightforward to code (e.g. for computer processing), and do not discriminate unduly on the basis of how articulate respondents are. On the other hand, they do not enable respondents to add any remarks, qualifications and explanations to the categories, and there is a risk that the categories might not be exhaustive and that there might be bias in them (Oppenheim, 1992, p. 115). Further, they can encourage mindless or less thought-through responses (Krosnick and Presser, 2010).

We consider in more detail below the different kinds of closed questions.

Scales of data

The questionnaire designer must choose the metric – the scale of data – to be adopted (Abascal and Diaz de Rada, 2014), and this will affect the possible statistical analysis. This concerns numerical data and which statistics can be used with which types of numerical data, and we advise readers to turn to Part 5 for an overview of the different scales of data that can be gathered (nominal, ordinal, interval and ratio), and the different statistics that can be used for analysis with them. Nominal data indicate categories; ordinal data indicate order ('high' to 'low', 'first' to 'last', 'smallest' to 'largest', 'strongly disagree' to 'strongly agree', 'not at all' to 'a very great deal'); ratio data indicate continuous values and a true zero (e.g. marks in a test, number of attendance per year, hours spent on study), thus:

QUESTION TYPE	*LEVEL OF DATA*
Dichotomous questions	Nominal
Multiple choice questions	Nominal
Rank ordering	Ordinal
Rating scales	Ordinal
Constant sum questions	Ordinal
Ratio data questions	Ratio
Open-ended questions	Word-based data

Dichotomous questions

A highly structured questionnaire asks closed questions. These can take several forms. *Dichotomous* questions have a 'yes'/'no' response, for example, 'have you ever had to appear in court?', 'do you prefer didactic methods to child-centred methods?'. The layout of a dichotomous question can be thus:

Sex (please tick): Male ☐ Female ☐

The dichotomous question is useful, for it compels respondents to 'come off the fence' on an issue. It provides for a clear, unequivocal response. Further, it is possible to code responses quickly, there being only two categories of response. A dichotomous question is also useful as a funnelling or sorting device for subsequent questions, for example: 'If you answered "yes" to question X, please go to question Y; if you answered "no" to question X, please go to question Z' (see the section below on contingency, skip and branching questions). This applies to paper-based questionnaires. In electronic/Internet questionnaires, based on the responses given, the computer can automatically take the respondent directly to the appropriate next place in the questionnaire, without instructions being given. Sudman and Bradburn (1982, p. 89) suggest that if dichotomous questions are being used, then it is desirable to use several to gain data on the same topic, in order to reduce the problems of respondents 'guessing' answers.

On the other hand, the researcher must ask whether a 'yes'/'no' response actually provides any useful information. Requiring respondents to make a 'yes'/'no' decision may be inappropriate; it might be more appropriate to have a range of responses, for example in a rating scale. A 'yes' or a 'no' may be inappropriate for a situation whose complexity is better served by a series of questions which catch that complexity. Further, Youngman (1984, p. 163) suggests that it is a natural human tendency to agree rather than to disagree with a statement ('acquiescence', discussed below); this suggests that a simple dichotomous question might build in respondent bias. People may be more reluctant to agree with a negative statement than to disagree with a positive question (Weems *et al*., 2003).

In addition to dichotomous questions ('yes'/'no' questions), a piece of research might ask for information about further dichotomous variables, for example, gender (male/female), type of school (elementary/secondary), type of course (vocational/non-vocational). Here, again, only one of two responses can be selected. Such nominal data can then be processed using the chi-square statistic, the binomial test, the G-test and cross-tabulations (for examples, see Cohen and Holliday, 1996). Dichotomous questions are treated as nominal data (see Part 5).

Multiple-choice questions

To try to gain some purchase on complexity, the researcher can move towards *multiple-choice* questions, where the range of choices is designed to include the likely range of responses to given statements. Champagne (2014) argues against the use of residual categories such as 'other', as these might insult the respondent, suggesting that the researcher has not done sufficient preparation work in identifying the likely categories of response, i.e. it is important to avoid items which have 'missing choices' (p. 41).

For example, the researcher might ask a series of questions about a new chemistry scheme in the school; a statement precedes a set of responses thus:

The New Intermediate Chemistry Education (NICE) is:

(a) a waste of time;
(b) an extra burden on teachers;
(c) not appropriate to our school;
(d) a useful complementary scheme;
(e) a useful core scheme throughout the school;
(f) well-presented and practicable.

The categories have to be discrete (i.e. having no overlap, being mutually exclusive) and have to exhaust the possible range of responses. Guidance has to be given on the completion of the multiple-choice, clarifying, for example, whether respondents are able to tick only *one* response (a *single answer* mode) or a *constrained number* of choices (e.g. three priorities from a list of ten possible choices) or a *free choice* (tick as many as you wish from the list). Like dichotomous questions, multiple-choice questions can be quickly coded and quickly aggregated to give frequencies of response. If that is appropriate for the research, then this might be a useful instrument.

The layout of a multiple-choice question can be thus:

Number of years in teaching
1–5 ☐ 6–14 ☐ 15–24 ☐ 25+ ☐

Which age group do you teach at present (you may tick more than one)?
Infant/kindergarten ☐
Primary ☐
Secondary (excluding sixth form) ☐
Sixth form only ☐

Just as dichotomous questions have their parallel in dichotomous variables, so multiple-choice questions have their parallel in *multiple elements of a variable*. For example, the researcher may be asking to which form a student belongs – there being up to, say, forty forms in a large school, or the researcher may be asking which post-16 course a student is following (e.g. academic, vocational, interest-based). In these cases only one response may be selected. As with the dichotomous variable, the listing of several categories or elements of a variable (e.g. form membership and course followed) enables nominal data to be collected and processed using the chi-square statistic, the G-test and cross-tabulations (Cohen and Holliday, 1996). Multiple-choice questions are treated as nominal data (see Part 5).

It is important to include in the multiple choices those that will enable respondents to select the response that most closely represents their view; hence a pilot is needed to ensure that the categories are comprehensive, exhaustive and representative. On the other hand, the researcher may be interested only in certain features, and it is these which would figure in the response categories.

The multiple-choice questionnaire seldom gives more than a crude statistic, for words are inherently ambiguous. In the example above, of chemistry, the notion of 'useful' is unclear, as are 'appropriate', 'practicable' and 'burden'. Respondents could interpret these words differently in their own contexts, thereby rendering the data ambiguous. One respondent might see the utility of the chemistry scheme in one area and thereby say that it is useful – ticking category (d). Another respondent might see the same utility in that same one area, but, because it is only useful in that single area, may see this as a flaw and therefore not tick category (d). With an anonymous questionnaire this difference is impossible to detect.

This is the heart of the problem of questionnaires: different respondents interpret the same words differently. 'Anchor statements' can be provided to allow a degree of discrimination in response (e.g. 'strongly agree', 'agree' etc.) and this yields greater reliability (Krosnick and Presser, 2010), but there is still no guarantee that respondents will interpret them in the way that is intended. In the example above this might not be a problem as the researcher might only be seeking an index of utility, without wishing to know the areas of utility or the reasons for that utility. The evaluator might be wishing only for a crude statistic (which might be very useful in making a decisive judgement about a programme). In this case a rough and ready statistic might be perfectly acceptable.

One can see in the example of chemistry above not only ambiguity in the wording but a very incomplete set of response categories which is hardly capable of including all aspects of the chemistry scheme. That this might be politically expedient cannot be overlooked, and if the choice of responses is limited then those responses might build bias into the research. For example, if the responses were limited to statements about the *utility* of the chemistry scheme, then the evaluator would have little difficulty in establishing that the scheme was useful. By avoiding the inclusion of negative statements or the opportunity to record a negative response the research will surely be biased.

Multiple-choice items are also prone to problems of word order and statement order. For example, Dillman *et al*. (2003, p. 6) report a study of sports, in which tennis was found to be less exciting than football when the tennis option was presented before the football option, and more exciting when the football option was placed before the tennis option. This suggests that respondents tend to judge later items in terms of the earlier items, rather than vice versa, and that they overlook features specific to later items if these are not contained in the earlier items. This is an instance of the 'primacy effect' or 'order effect', wherein items earlier in a list are given greater weight than items lower in the list. Order effects are resilient to efforts to minimize them, and primacy effects are particularly strong in Internet questionnaires (p. 22). Preceding questions and the answers given may influence responses to subsequent questions (Schwartz *et al*., 1998, p. 177).

Order effects and the primacy effect are examples of context effects, in which some questions (sometimes coming later in the questionnaire, as respondents do not always answer questions in the given sequence, and may scan the whole questionnaire before answering specific items) may affect the responses given to other questions (Friedman and Amoo, 1999, p. 122), biasing the responses by creating a specific mindset, i.e. a predisposition to answering questions in a particular way.

Further, questionnaires designers must be aware of the recency effect, i.e. respondents tend to remember the last item in a list rather than what precedes it, and this affects their response.

Rank ordering

The rank order question is akin to the multiple-choice question in that it identifies options from which respondents can choose, yet it moves beyond multiple-choice items in that it asks respondents to identify priorities. This enables a *relative* degree of preference, priority, intensity etc. to be charted. Rank ordering requires respondents to *compare* values across variables; in this

respect they are unlike rating scales in which the values are entered independently of each other (Ovadia, 2004, p. 404), i.e. the category 'strongly agree' can be applied to a single variable without any regard to what one enters for any other variable. In a ranking exercise the respondent is required to take account of the other variables, because he/she is being asked to see their *relative* value, weighting or importance. This means that, in a ranking exercise, the task must be fair, i.e. the variables are truly able to be compared and placed in a rank order, and they lie on the same scale and/or can be judged on the same criteria.

In the rank ordering exercise, a list of factors is set out and the respondent is required to place them in a rank order, for example:

Please indicate your priorities by placing numbers in the boxes to indicate the ordering of your views, 1 = the highest priority, 2 = the second highest, and so on.

The proposed amendments to the mathematics scheme might be successful if the following factors are addressed:

- the appropriate material resources are in school; ☐
- the amendments are made clear to all teachers; ☐
- the amendments are supported by the mathematics team; ☐
- the necessary staff development is assured; ☐
- there are subsequent improvements to student achievement; ☐
- the proposals have the agreement of all teachers; ☐
- they improve student motivation; ☐
- parents approve of the amendments; ☐
- they will raise the achievements of the brighter students; ☐
- the work becomes more geared to problem-solving. ☐

In this example ten items are listed. Whilst this might be enticing for the researcher, enabling fine distinctions possibly to be made in priorities, it might be asking too much of the respondents to make such distinctions. They genuinely might not be able to differentiate their responses, or they simply might not feel strongly enough to make such distinctions. The inclusion of too long a list might be overwhelming. Indeed Wilson and McLean (1994, p. 26) suggest that it is unrealistic to ask respondents to arrange priorities where more than five ranks are requested. In the case of the list of ten points above, the researcher might approach this problem in one of two ways. The list in the questionnaire item can be reduced to five items only, in which case the *range* and comprehensiveness of responses that fairly catches what the respondent feels is significantly reduced. Alternatively, the list of ten items can be retained, but the request can be made to the respondents only to rank their first five priorities, in which case the range is retained and the task is not overwhelming (though the problem of sorting the data for analysis is increased).

An example of a shorter list might be:

Please place these in rank order of the most to the least important, by putting the position (1–5) against each of the following statements, number one being the most important and number 5 being the least important:

Students should enjoy school	[ ]
Teachers should set less homework	[ ]
Students should have more choice of subjects in school	[ ]
Teachers should use more collaborative methods	[ ]
Students should be tested more, so that they work harder	[ ]

Rankings may also assume that the different items can truly be placed on a single scale. Consider the example above, where the respondent is required to place five items on a single scale of importance. Can these items really be differentiated according to the single criterion of 'importance'? Surely 'fitness for purpose' and context would suggest that a fairer answer is that 'it all depends' on what is happening in a specific context, i.e. even though one could place items in a rank order, in fact it may be meaningless to do so. The items may truly not be comparable (Ovadia, 2004, p. 405). As Ovadia (2004, p. 407) notes, valuing justice may say nothing about valuing love, so to place them in a single ranking scale of importance may be meaningless.

Rankings are useful in indicating *degrees* of response. In this respect they are like rating scales, discussed below. Ranking questions are treated as ordinal data (see Part 5 for a discussion of ordinal data). However, rankings do not enable sophisticated statistical analysis to be conducted (Ovadia, 2004, p. 405), as the ranks are inter-dependent rather than independent, and these vary for each respondent, i.e. not only does the rank '1st' mean different things to different respondents, but there are no equal intervals between each rank, and the rank of, say, '3rd' has a different

meaning for different respondents, which is relative to their idea of what constitutes '2nd' and '4th', i.e. the rankings are inter-dependent; there is no truly common metric here. Further, because rankings force a respondent to place items in a rank order, differences between values may be overstated. The difference between ranks 1 and 2 might be large, whereas the difference between ranks 5 and 6 might be negligible; simply placing items in a ranks order here, therefore, might be dangerous if too much weight is put on them.

Rankings operate on a zero-sum model (Ovadia, 2004, p. 406), i.e. if one places an item in first position then this means that another item drops in the ranking; this may or may not be desirable, depending on what the researcher wishes to find out. Researchers using rankings will need to consider whether it is fair to ask respondents really to compare items and to judge one item in relation to another; to ask 'are they really commensurable?' (able to be measured by the same single standard or criterion). It might be preferable to use rating scales.

Rating scales

One way in which degrees of response, intensity of response and the move away from dichotomous questions and rankings have been managed can be seen in the notion of *rating scales*: Likert scales, semantic differential scales, Thurstone scales and Guttman scaling. These are useful devices for the researcher, as they build in a degree of sensitivity and differentiation of response whilst still generating numbers. Here we focus on the first two of these, though readers will find the others discussed in Oppenheim (1992), Krosnick and Presser (2010) and Dillman *et al*. (2014). A Likert scale (named after its deviser, Rensis Likert, 1932) provides a range of responses to a given question or statement, for example:

How important do you consider work placements to be for secondary school students?

1 = not at all
2 = very little
3 = a little
4 = quite a lot
5 = a very great deal

All students should have access to free higher education.

1 = strongly disagree
2 = disagree
3 = neither agree nor disagree
4 = agree
5 = strongly agree

Such a scale could be set out thus:

Please complete the following by placing a tick in one space only, as follows:

1 = strongly disagree; 2 = disagree;
3 = neither agree nor disagree;
4 = agree; 5 = strongly agree

Senior school staff should teach more

1	2	3	4	5
[ ]	[ ]	[ ]	[ ]	[ ]

In these examples the categories need to be discrete and to exhaust the range of possible responses which respondents may wish to give. Notwithstanding the problems of interpretation which arise as in the previous example – one respondent's 'agree' may be another's 'strongly agree', one respondent's 'very little' might be another's 'a little' – the greater subtlety of response which is built into a rating scale renders this a very attractive and widely used instrument in research, particularly for gathering data on attitudes and opinions.

These two examples both indicate an important feature of an attitude scaling instrument, namely the assumption of *unidimensionality* in the scale; the scale should only be measuring one thing at a time (Oppenheim, 1992, pp. 187–8). Indeed this is a cornerstone of Likert's own thinking (1932).

It is a very straightforward matter to convert a dichotomous question into a multiple-choice question. For example, instead of asking dichotomous ('yes/no') questions such as 'do you?', 'have you?', 'are you?', 'can you?', a simple addition to wording will convert it into a much more subtle rating scale, by substituting the words 'to what extent?', 'how far?', 'how much?', 'how often?' etc.

A semantic differential is a variation of a rating scale which operates by putting an adjective at one end of a scale and its opposite at the other, for example:

How informative do you consider the new set of history textbooks to be?

	1	2	3	4	5	6	7	
useful	_	_	_	_	_	_	_	useless

Respondents indicate their opinion by circling or putting a mark on that position on the scale which most represents what they feel. Researchers devise their own terms and their polar opposites, for example:

Approachable	… Unapproachable
Generous	… Mean
Friendly	… Hostile
Caring	… Uncaring
Attentive	… Inattentive
Hard-working	… Lazy

Osgood *et al.* (1957), the pioneers of this technique, suggest that semantic differential scales are useful in three contexts: *evaluative* (e.g. valuable–valueless, useful–useless, good–bad); *potency* (e.g. large–small, weak–strong, light–heavy); and *activity* (e.g. quick–slow, active–passive, dynamic–lethargic). However, Champagne (2014) argues against not defining each scale point, as, if numbers have no anchor statement indicating what each point means, respondents people will interpret them very differently, building in unreliability. He strongly advocates including descriptors for every scale point, as this makes for greater clarity, reduced ambiguity and more useable results, for example: 1 = definitely no; 2 = probably no; 3 = probably yes; 4 = definitely yes (p. 45).

There are several commonly used categories in rating scales, for example:

- strongly disagree/disagree/neither agree nor disagree/agree/strongly agree
- very seldom/occasionally/quite often/very often
- very little/a little/somewhat/quite a lot/a very great deal
- never/almost never/sometimes/often/very often
- not at all important/unimportant/neither important nor unimportant/important/very important
- very true of me/a little bit true of me/don't know/not really true of me/very untrue of me
- strongly agree/agree/uncertain (or 'neither agree nor disagree')/disagree/strongly agree

To these could be added the category 'don't know' or 'have no opinion'. On the other hand, Krosnick and Presser (2010, p. 284) and Champagne (2014) warn that the inclusion of a 'don't know' category can be used by respondents not as a genuine category but because of satisficing (discussed below), intimidation and self-protection, ambivalence and problems in understanding the question or how to respond. Indeed Krosnick and Presser note (p. 285) that the inclusion of this category might compromise the quality of the data, and this might apply particularly if sensitive questions are being asked where socially undesirable response categories are included (p. 287).

Champagne (2014) reminds researchers of the need to ensure that the response scale actually matches the item (p. 47). For example, many researchers will use categories such as: 'strongly agree', 'agree', 'neither agree nor disagree', 'disagree' and 'strongly disagree', but the questions to which these scale points are used might be inappropriate, for example:

- 'I regularly spoke with my department head' (which concerns consistency and frequency);
- 'I meet my department head by appointment' (which requires either a 'yes/no' answer or an answer concerning frequency);
- 'The pace of this staff meeting was satisfactory' (which does not tell us, for example, whether the pace was fast or slow, or too fast or too slow).

Rating scales are widely used in research, and rightly so, for they combine the opportunity for a flexible response with the ability to determine frequencies, correlations and other forms of quantitative analysis. They afford the researcher the freedom to fuse measurement with opinion, quantity and quality.

Though rating scales are useful in research, the investigator, nevertheless, needs to be aware of their limitations. For example, the researcher may infer a degree of sensitivity and subtlety from the data that they cannot bear. There are other cautionary factors about rating scales and we set these out below.

1 Equal intervals

There is no assumption of equal intervals between the categories, hence a rating of 4 indicates neither that it is twice as powerful as 2 nor that it is twice as strongly felt; one cannot infer that the intensity of feeling in the Likert scale between 'strongly agree' and 'agree' somehow matches the intensity of feeling between 'strongly disagree' and 'disagree'. These are illegitimate inferences. The problem of equal intervals has been addressed in Thurstone scales (Thurstone and Chave, 1929; Oppenheim, 1992, pp. 190–5). Friedman and Amoo (1999, p. 115) suggest that if the researcher wishes to assume equal intervals ('equal-sized gradations') between points in the rating scale, then he or she must ensure that the category descriptors are genuinely equal interval. Take, for example, the scale 'not at all', 'very little', 'a little', 'quite a lot', 'a very great deal'. Here the conceptual distance between 'a little' and 'quite a lot' is much greater than between 'very little' and 'a little', i.e. there are not equal intervals.

2 The meaning of numbers

Numbers have different meanings for different respondents, so one person may use a particular criterion to award a score of '6' on a seven-point scale, whilst

another person using exactly the same criterion would award a score of '5' on the same scale. Here '6' and '5' actually mean the same but the numbers are different. Alternatively, one person looking at a score of, say, 7 marks out of 10 on a ten-point scale would consider that to be a high score, whereas another person looking at the same score would consider it to be moderate only. Similarly, the same word has a different meaning for different respondents; one teacher may think that 'very poor' is a very negative descriptor, whereas another might think less negatively about it, and what one respondent might term 'poor', another respondent, using the same criterion, might term 'very poor'. Friedman and Amoo (1999, p. 115) report that there was greater consistency between subjects on the meanings of positive words rather than negative words, and they suggest that, therefore, researchers should use descriptors that have lesser strength at the negative pole of a scale (p. 3). Further, they suggest that temporal words (e.g. 'very often', 'seldom', 'fairly often', 'occasionally' etc.) are open to great variation in their meanings for respondents (p. 3).

3 Unrealistic choices

Some rating scales are unbalanced, forcing unrealistic choices to be made, for example, in the scale 'very acceptable', 'quite acceptable', 'a little acceptable' 'acceptable' and 'unacceptable', or in the scale 'excellent, 'very good', 'quite good', 'good' and 'poor', there are four positive categories and only one negative category (cf. Friedman and Amoo, 1999, p. 119). This can skew results. Such imbalance could even be called unethical.

4 Layout effects

Respondents are biased towards the left-hand side of a bipolar scale (Friedman and Amoo, 1999, p. 120; Hartley and Betts, 2010, p. 25). For example, if the scale 'extremely good' to 'extremely poor' runs from left to right respectively, then the results will be different when the same scale is reversed ('extremely poor' to 'extremely good') and runs from left to right (or, for example, 'strongly agree' on the left to 'strongly disagree' on the right, and vice versa). Typically, categories on the left-hand side of a scale are used more frequently than those on the right-hand side of a scale. Hartley and Betts (2010, p. 25) found that those scales which had a positive label on the left-hand side would elicit higher scores than other orderings. Hence researchers must be cautious about putting all the positive categories on the left-hand side alone, as this can result in more respondents using those categories than if they were placed on the right-hand side of the scale, i.e. rating scales may want to mix the item scales so that sometimes there are positive scores on the left and sometimes positive scores on the right (but too much mixing is confusing and might lead to a non-response or an unintended response).

5 Direction of comparison

The 'direction of comparison' (Friedman and Amoo, 1999, p. 120) also makes a difference to results. The authors cite an example where students were asked how empathetic their male and female teachers were in regard to academic and personal problems. When the question asked 'would you say that female teachers were more empathetic … than the male teachers?' the mean score of the responses on a nine-point scale was different from when the question was 'would you say that male teachers were more empathetic … than the female teachers?' In the former, 41 per cent of responses indicated that female teachers were more empathetic, whereas in the latter only 9 per cent of responses indicated that female teachers were more empathetic.

6 Truthfulness of responses

We have no check on whether respondents are telling the truth. Some may be deliberately falsifying their replies.

7 Inadequate categories

We have no way of knowing if the respondent wishes to add any other comments about the issue under investigation. It might be that there is something far more pressing about the issue than the rating scale includes but which is condemned to silence for want of a category. A straightforward way to circumvent this issue is to run a pilot and also to include a category entitled 'other (please state)'.

8 Number of scale points

Most of us would not wish to be called extremists; we may prefer to appear like each other in many respects. For rating scales this means that we avoid the two extreme poles at each end of the continuum of the rating scales, reducing the number of positions in the scales to a choice of three (in a five-point scale). That means that *in fact* there could be very little choice for us. Further, a trichotomous scale (dislike/neutral/like) may not catch the sensitivity of a larger scale, for example, a respondent may wish to record a 'moderately like' response but there is no category for this. The way round these problems is to create a larger scale than a five-point scale, for example a seven-point scale. To go beyond a seven-point scale is to invite a

degree of detail and precision which might be inappropriate for the item in question, particularly if the argument set out above is accepted, namely, that one respondent's scale point 3 might be another's scale point 4.

Friedman and Amoo (1999, p. 120) suggest that five-point to eleven-point scales might be most useful, whilst Schwartz *et al*. (1991, p. 571) and Krosnick and Presser (2010) suggest that seven-point scales seem to be preferable in terms of reliability, the ability of respondents to discriminate between the values in the scales, and the percentages of respondents who are 'undecided'. If a differentiated, fine-grained response is sought then a larger (five- or seven-point) rather than a smaller scale is preferable. Indeed they suggest that reliability is lower for those scales which have few scale points and higher for those which have more scale points, levelling off from seven points or more, with validity lowering if there are many scale points.

9 Labelling scale points

Schwartz *et al*. (1991, p. 571), Krosnick and Presser (2010) and Champagne (2014) report that rating scales that have a verbal label for each point in the scale are more reliable than rating scales which provide labels only for the end-points of the numerical scales. Indeed Krosnick and Presser (2010) report that respondents prefer to have such labels.

10 Ratio data

If the researcher wishes to use ratio data (see Part 5) in order to calculate more sophisticated statistics (e.g. regressions, factor analysis, structural equation modelling), then a ratio scale must have a true zero ('0') and equal intervals. Many rating scales use an eleven-point scale here that runs from 0 to 10, with 0 being 'not at all' (or something equivalent to this, depending on the question/item) and 10 being the highest score (e.g. 'completely' or 'excellent').

11 End-point descriptors

The end-point descriptors on a scale have a significant effect on the responses (Friedman and Amoo, 1999, p. 117). For example, if the end-points of a scale are extreme (e.g. 'terrible' and 'marvellous') then respondents will avoid these extremes, whereas if the end-points are 'very bad' and 'very good' then more responses in these categories are chosen.

12 Number, nature and order of scale points

The nature of the scaling may affect significantly the responses given and the range of responses actually given (Schwartz and Bienias, 1990, p. 63). Hartley and Betts (2010) and Dillman *et al*. (2014) note that even different rating scale order exerts an influence on responses. Schwartz *et al*. (1991) found that if a scale only had positive integers (e.g. 1 to 10) on a scale of 'extremely successful' to 'not at all successful', then 34 per cent of respondents chose values in the 1–5 categories. However, when the scale was set at −5 for 'not at all successful' and +5 for 'extremely successful', then only 13 per cent of respondents chose the equivalent lower five values (−5 to 0). The authors surmised that the former scale (0–10) was perceived by respondents to indicate degrees of success, whereas the latter scale (−5 to 0) was perceived by respondents to indicate not only the absence of success but the presence of the negative factor of failure (see also Schwartz *et al*., 1998, p. 177). Indeed they reported that respondents were reluctant to use negative scores (1991, p. 572) and that responses to a −5 to +5 scale tended to be more extreme than responses to a 0–10 scale, even when they used the same scale verbal labels.

Schwartz *et al*. also suggest (1991, p. 577) that, in a −5 to +5 scale, zero (0) indicates absence of an attribute, whereas in a 0–10 scale a zero (0) indicates the presence of the negative end of the bipolar scale, i.e. the zero has two different meanings, depending on the scale used. Hence researchers must be careful about not only the verbal labels that they use, but also the scales and scale points that they use with those same descriptors. Kenett (2006, p. 409) also comments, in this respect, that researchers will need to consider whether they are asking about a bipolar dimension (e.g. 'very successful' to 'very unsuccessful') where an attribute and its opposite are included, or whether a single pole is being used (e.g. only degrees of positive response or presence of a factor). For a bipolar dimension, a combination of negative and positive numbers on a scale may be useful (with the cautions indicated above), whereas for a single polar dimension then only positive numbers should be used (cf. Schwartz *et al*., 1991, p. 577). In other words, if the researcher is looking to discover the intensity of a single attribute then it is better to use positive numbers only (p. 578).

13 Terminology of response categories

Response alternatives may signal the nature of the considerations to be borne in mind by respondents (Gaskell *et al*., 1994, p. 243). For example, if one is asking about how often there are incidents of indiscipline in a class, the categories 'several times each lesson', 'several times each morning', 'several times each day' may indicate that a more inclusive, wider definition of 'indiscipline' is required than if the categories of 'several times each week', 'several times each month'

or 'several times each term' were used. The terms used may frame the nature of the thinking or responses that the respondent uses. Gaskell *et al*. suggest that this is particularly the case if some vague phrases are included in the response categories (p. 242). Obtained responses, as Schwartz and Bienias (1990, p. 62) indicate, are a function of the response alternatives that the researcher has provided. Indeed Bless *et al*. (1992, p. 309) indicate that scales which offer higher response categories/values tend to produce higher estimates from the respondents and that this tendency increases as questions become increasingly difficult (p. 312).

14 Clustering of responses

Respondents tend to cluster their responses (e.g. around the centre or around one end or another of the scale), and their responses to one item may affect their responses to another item (i.e. creating a single mindset).

15 Forced choices

Choices may be 'forced' by omitting certain categories (e.g. 'no opinion', 'undecided', 'don't know', 'neither agree nor disagree'). If the researcher genuinely believes that respondents do, or should, have an opinion then such omissions may be justified. Alternatively, it may be unacceptable to force a choice for want of a category which genuinely lets respondents say what is in their minds, even if their minds are not made up about a factor or if they have a reason for concealing their true feelings. Forcing a choice may lead to respondents having an opinion on matters that they really have no opinion about (Friedman and Amoo, 1999, p. 118), and respondents may object to them.

16 Mid-points

There is a tendency for participants to opt for the mid-point of a 5- or seven-point scale (the central tendency). This is notably an issue in East Asian respondents, where the 'doctrine of the mean' is advocated in Confucian culture. One way to overcome this is to use an even number scaling system, as there is no mid-point. On the other hand, it could be argued that if respondents wish to 'sit on the fence' and choose a mid-point, then they should be given the option to do so. Krosnick and Presser (2010, p. 274) note that reliability and validity increase with the use of mid-points, and they advise using them.

On some scales there are mid-points; on the five-point scale it is category 3, and on the seven-point scale it is category 4. However, choosing an even number of scale points, for example a six-point scale, might *require* a decision on rating to be indicated. For example, suppose a new staffing structure has been introduced into a school and the headteacher/principal is seeking some guidance on its effectiveness. A six-point rating scale might ask respondents to indicate their response to the statement:

> The new staffing structure in the school has enabled teamwork to be managed within a clear model of line management.
>
> (Circle one number)
>
> | | 1 | 2 | 3 | 4 | 5 | 6 | |
> |---|---|---|---|---|---|---|---|
> | strongly agree | _ | _ | _ | _ | _ | _ | strongly disagree |

Let us say that one member of staff circled 1, eight staff circled 2, twelve staff circled 3, nine staff circled 4, two staff circled 5 and seven staff circled 6. There being no mid-point on this continuum, the researcher could infer that those respondents who circled 1, 2 or 3 were in some measure of agreement, whilst those respondents who circled 4, 5 or 6 were in some measure of disagreement. That would be very useful for, say, a headteacher/principal in publicly displaying agreement, there being twenty-one staff (1+8+12) agreeing with the statement and eighteen (9+2+7) displaying a measure of disagreement. However, one could point out that the measure of 'strongly disagree' attracted seven staff – a very strong feeling – which was not true for the 'strongly agree' category, which attracted only one member of staff. The extremity of the voting has been lost in a crude aggregation.

Further, if the researcher were to aggregate the scoring around the two mid-point categories (3 and 4), there would be twenty-one members of staff represented, leaving nine (1+8) from categories 1 and 2 and nine (2+7) from categories 5 and 6; adding together categories 1, 2, 5 and 6, a total of eighteen is reached, which is less than the twenty-one total of the two categories 3 and 4. It seems on this scenario that it is far from clear that there was agreement with the statement from the staff; indeed taking the high incidence of 'strongly disagree', it could be argued that those staff who were perhaps ambivalent (categories 3 and 4), coupled with those who registered a 'strongly disagree', indicate not agreement but disagreement with the statement.

The interpretation of data has to be handled very carefully; ordering them to suit a researcher's own purposes might be very alluring but quite illegitimate. The golden rule here is that crude data can only yield crude interpretation; subtle interpretations require subtle data. The interpretation of data must not distort the data unfairly. Rating scale questions are treated as ordinal

data (see Part 5) and use modal scores and non-parametric data analysis, though one can find very many examples where this rule has been violated, and non-parametric data have been treated as parametric data. Indeed there is an argument that, in fact, a Likert scale with anchor statements (scale point labels) is really a nominal scale, though typically it is taken to be ordinal.

It has been suggested that the attraction of rating scales is that they provide more opportunity than dichotomous questions for rendering data more sensitive and responsive to respondents. This makes rating scales particularly useful for tapping attitudes, perceptions and opinions. The need for a pilot study to devise and refine categories, making them exhaustive and discrete, has been suggested as a necessary part of this type of data collection.

Rating scales are more sensitive than dichotomous scales. Nevertheless they are limited in their usefulness to researchers by their fixity of response caused by the need to select from a given choice. A questionnaire might be tailored even more to respondents by including *open-ended* questions to which they can reply in their own terms and own opinions. For further reviews of rating scales, we refer the reader to Hartley and Betts (2010).

Ranking or rating?

If the researcher wishes respondents to *compare* variables (items) and award scores for items in relation to each other, then rankings are suitable. If the researcher wishes respondents to give a response/score to variables (items) that are independent of the score awarded to any other variables (items), then ratings should be considered. In the latter, the score that one awards to one variable has no bearing or effect on the score that one awards to another. In practice, the results of many rating scales may enable the researcher to place items in a rank order (Ovadia, 2004, p. 405), but rating scales may also result in many variables having ties (the same score) in the values given, which may be coincidental or, indeed, the 'result of indifference' (p. 405) on the part of the respondent to the variable in question (e.g. respondents simply and quickly tick the middle box (e.g. '3' in a five-point scale) going down a list of items).

Rankings may force the respondent to use the full range of the scale (the scale here being the number of items included, e.g. if there are ten items then up to ten rankings might be required). By contrast, ratings do not have such a stringent requirement; respondents may cluster their responses to all the items around one end of a scale (e.g. points '5', '6' and '7' in a seven-point scale, or point '3' in a five-point scale).

Let us imagine that a researcher asked respondents to indicate the importance of three items in respect of student success, and that the scale used was to award points out of ten. Here are the results for respondent A and respondent B (cf. Ovadia, 2004, p. 407):

Respondent A: working hard (9 points); family pressure (6 points); enjoyment of the subject (5 points).

Respondent B: working hard (6 points); family pressure (4 points); enjoyment of the subject (2 points).

A ranking exercise would accord the same positioning of the items on these two scores: in first place comes 'working hard', then 'family pressure', and in the lowest position, 'enjoyment of the subject'. However, as we can see, the *actual* scores are very different, and respondent A awards much higher scores than respondent B, i.e. for respondent A these items are much more important than for respondent B, and any single item is much more important for respondent A than for respondent B. Whilst rankings and ratings here yield equally valid results, the issue is one of 'fitness for purpose': if the researcher wishes to *compare* then rankings might be useful, whereas if the researcher wishes to examine actual values then ratings might be more useful.

Further, let us imagine that for respondent A in this example, the score for 'working hard' drops by 2 points over time, the score for 'family pressure' drops by 1 point and the score for 'enjoyment of the subject' drops by 3 points over time. The result of the ranking, however, remains the same, i.e. even though the level of importance has dropped for these three items, the ranking is insensitive to these changes.

Constant sum questions

In this type of question respondents are asked to distribute a given number of marks (points) between a range of items, for example:

'Please distribute a total of ten points among the sentences that you think most closely describe your behaviour. You may distribute these freely: they may be spread out, or awarded to only a few statements, or all allocated to a single sentence if you wish.'

I can take advantage of new opportunities	[ ]
I can work effectively with all kinds of people	[ ]
Generating new ideas is one of my strengths	[ ]
I can usually tell what is likely to work in practice	[ ]
I am able to see tasks through to the very end	[ ]
I am prepared to be unpopular for the good of the school	[ ]

This enables priorities to be identified, comparing highs and lows, and for equality of choices to be indicated, and, importantly, for this to be done in the respondents' own terms. It requires respondents to make comparative judgements and choices across a range of items. For example, we may wish to distribute ten points for aspects of an individual's personality:

Talkative	[ ]
Cooperative	[ ]
Hard-working	[ ]
Lazy	[ ]
Motivated	[ ]
Attentive	[ ]

This means that the respondent has to consider the *relative* weight of each of the given aspects before coming to a decision about how to award the marks. To accomplish this means that the all-round nature of the person, in the terms provided, has to be considered in order to see, on balance, which aspect is stronger when compared to another.

The difficulty with this approach is to decide how many marks can be distributed (a round number, e.g. ten, makes subsequent calculation easily comprehensible) and how many statements/items to include, for example, whether to have the same number of statements as there are marks, or more or fewer statements than the total of marks. Having too few statements/items does not do justice to the complexity of the issue, and having too many statements/items may mean that it is difficult for respondents to decide how to distribute their marks. Having too few marks available may be unhelpful, but, by contrast, having too many marks and too many statements/items can lead to simple computational errors by respondents. Our advice is to keep the number of marks to ten and the number of statements to around six to eight. Constant sum data are ordinal, and this means that non-parametric analysis can be performed on the data (see Part 5). Constant sum questions may place too great an onus on participants to decide, and this could lead to non-response or withdrawal.

Ratio data questions

We discuss ratio data in Part 5 and we refer the reader to the discussion and definition there. For our purposes here we suggest that ratio data questions deal with continuous variables where there is a true zero, for example:

How much money do you have in the bank?	____
How many times have you been late for school?	____
How many marks did you score in the mathematics test?	____
How old are you (in years)?	____

Here no fixed answer or category is provided, and the respondent puts in the numerical answer that fits his/her exact figure, i.e. the accuracy is higher, much higher than in *categories* of data. This enables averages (means), standard deviations, range and high-level statistics to be calculated, for example, regression, factor analysis, structural equation modelling (see Part 5).

An alternative form of ratio scaling is where the respondent has to award marks out of, say, ten, or a percentage, for a particular item (e.g. Kgaile and Morrison, 2006), as illustrated in Table 24.2.

This kind of scaling is often used in telephone interviews, as it is easy for respondents to understand. The argument could be advanced that this is a sophisticated form of rating scale, but the terminology used in the instruction clearly suggests that it asks for ratio scale data. Ratio data that ask for a percentage assume a degree of sensitivity that may be unwarranted, i.e. what

TABLE 24.2 A MARKING SCALE IN A QUESTIONNAIRE

'Please give a mark from 0 to 10 for the following statements, with 10 being excellent and 0 being very poor. Please circle the appropriate number for each statement.'

Teaching and learning	*Very poor*										*Excellent*
1 The attention given to teaching and learning at the school	0	1	2	3	4	5	6	7	8	9	10
2 The quality of the lesson preparation	0	1	2	3	4	5	6	7	8	9	10
3 How well learners are cared for, guided and supported	0	1	2	3	4	5	6	7	8	9	10
4 How effectively teachers challenge and engage learners	0	1	2	3	4	5	6	7	8	9	10
5 The educators' use of assessment for maximizing learners' learning	0	1	2	3	4	5	6	7	8	9	10
6 How well students apply themselves to learning	0	1	2	3	4	5	6	7	8	9	10
7 Discussion and review by educators of the quality of teaching and learning	0	1	2	3	4	5	6	7	8	9	10

is the real difference between 70 per cent and 71 per cent or between 31 per cent and 32 per cent?

Champagne (2014) advises against grouping what should be ratio data into categorical data (groups). He gives the example of asking how many scoops of ice cream a person would like (p. 54): we would not expect a person to say 'between 1 and 4'; rather we would expect an exact answer. The person who answers '4 scoops' could well be a glutton, whilst the person who answers '1 scoop' could well be more diet conscious, but this point would be lost if the data were put into a group of '1–4'. If categories/groups are to be used, they must be meaningful and reasonable, i.e. 'all the choices within the group are treated as equal', so that within-group differences are unimportant and between-group differences are realistic (p. 56). This applies, for example, if we are grouping the ages of people into age groups. Champagne (p. 103) argues for keeping the ratio scale (and indeed the rating scale) as it is rather than regrouping/combining them into categories.

Matrix questions

Matrix questions do not concern types of questions but their layout, enabling the same kind of response to be given to several questions, for example 'strongly disagree' to 'strongly agree'. The matrix layout helps to save space, for example:

Please complete the following by placing a tick in one space only, as follows:

1 = not at all; 2 = very little; 3 = a moderate amount; 4 = quite a lot; 5 = a very great deal

How much do you use the following for assessment purposes?

		1	2	3	4	5
a	commercially published tests	[ ]	[ ]	[ ]	[ ]	[ ]
b	your own made-up tests	[ ]	[ ]	[ ]	[ ]	[ ]
c	students' projects	[ ]	[ ]	[ ]	[ ]	[ ]
d	essays	[ ]	[ ]	[ ]	[ ]	[ ]
e	samples of students' work	[ ]	[ ]	[ ]	[ ]	[ ]

Here five questions have been asked in only five lines, excluding, of course, the instructions and explanations of the anchor statements. Such a layout is economical on space.

A second example indicates how a matrix design can save a considerable amount of space in a questionnaire. Here the size of potential problems in conducting a piece of research is asked for, and data on how much these problems were soluble are requested. For the first issue (the size of the problem), 1 = no problem, 2 = a small problem, 3 = a moderate problem, 4 = a large problem, 5 = a very large problem. For the second issue (how much the problem was solved), 1 = not solved at all, 2 = solved only a very little, 3 = solved a moderate amount, 4 = solved a lot, 5 = completely solved (see Table 24.3).

Here thirty questions (15 × 2) have been covered in just a short amount of space.

Laying out the questionnaire like this enables the respondent to fill in the questionnaire rapidly. On the other hand, it risks creating a mindset in the respondent (a 'response set' (Baker, 1994, p. 181)) in that the respondent may simply go down the questionnaire columns and write the same number each time (e.g. all number 3) or, in a rating scale, tick all number 3. Such response sets can be detected by looking at patterns of replies and eliminating response sets from subsequent analysis (though this may be illegitimate, as the researcher has no way of knowing what was in the respondent's mind).

The conventional way of minimizing response sets has been by reversing the meaning of some of the questions so that the respondents will need to read them carefully. However, Weems *et al*. (2003) argue that using positively and negatively worded items within a scale is not measuring the same underlying traits. They report that some respondents will tend to disagree with a negatively worded item, that the reliability levels of negatively worded items are lower than for positively worded items and that negatively worded items receive greater non-response than positively worded items. Indeed they argue against mixed-item formats, and supplement this by reporting that inappropriately worded items can induce an artificially extreme response which, in turn, compromises the reliability of the data. Mixing negatively and positively worded items in the same scale, they argue, compromises both validity and reliability. Indeed they suggest that respondents may not read negatively worded items as carefully as positively worded items. Mixing positively and negatively worded items is confusing for respondents.

Contingency and skip questions, filters and branches

Contingency and skip questions depend on responses to earlier questions, for example: 'if your answer to question (1) was 'yes' please go to question (4)'. The earlier question acts as a filter for the later question, and the later question is contingent on the earlier, and is a branch of the earlier question. Some questionnaires will spell out the number of the question to which to go (e.g. 'please go to question 6'); others (in paper versions) will place an arrow to indicate the next question to be

TABLE 24.3 POTENTIAL PROBLEMS IN CONDUCTING RESEARCH

Potential problems in conducting research	*Size of the problem (1–5)*	*How much the problem was solved (1–5)*
1 Gaining access to schools and teachers; 2 Gaining permission to conduct the research (e.g. from principals); 3 Resentment by principals; 4 People vetting what could be used; 5 Finding enough willing participants for your sample; 6 Schools suffering from 'too much research' by outsiders and insiders; 7 Schools/people not wishing to divulge information about themselves; 8 Schools not wishing to be identifiable, even with protections guaranteed; 9 Local political factors that impinge on the school; 10 Teachers' fear of being identified/traceable, even with protections guaranteed; 11 Fear of participation by teachers (e.g. if they are critical of the school or others they could lose their contracts); 12 Unwillingness of teachers to be involved because of their workload; 13 The principal deciding on whether to involve the staff, without consultation with the staff; 14 Schools'/institutions' fear of criticism/loss of face; 15 The sensitivity of the research: the issues being investigated.		

answered if your answer to the first question was such-and-such.

A funnelling process moves from the general to the specific, asking questions about the general context or issues and then moving towards specific points within that. A filter is used to include and exclude certain respondents, i.e. to decide if certain questions are relevant or irrelevant to them, and to instruct respondents about how to proceed (e.g. which items to jump to or proceed to). For example, if respondents indicate a 'yes' or a 'no' to a certain question, then this might exempt them from certain other questions in that section or subsequently.

Contingency and filter questions may be useful for the researcher, but, in a paper-based questionnaire, they can be confusing for the respondent as it is not always clear how to proceed through the sequence of questions and where to go once a particular branch has been completed. Redline *et al*. (2002) found that respondents tend to ignore, misread and incorrectly follow branching instructions, such that item non-response occurs for follow-up questions that are only applicable to certain sub-samples, and respondents skip over, and therefore fail to follow up on, those questions that they should have completed. The authors found that the increased complexity of the questionnaire brought about by branching instructions negatively influenced its correct completion.

Redline *et al*. report (2002, p. 7) that the number of words in the question affects the respondents' ability to follow branching instructions: the more words there are in the question, the greater is the likelihood of the respondents overlooking the branching instructions. They also report that up to seven items, and no more, can be retained in the short-term memory. This has implications for the number of items in a list of telephone interviews, where there is no visual recall or checking possible. Similarly, the greater the number of answer categories, the greater is the likelihood of making errors, for example, overlooking branching instructions (p. 19). They report that respondents tend to see branching instructions when they are placed by the last category, particularly if they have chosen that last category.

Redline *et al*. note (2002, p. 8) that sandwiching branching instructions between items which do not branch is likely to lead to errors of omission and commission being made: omitting to answer all the questions and answering the wrong questions respectively. Further, locating the instructions for branching some distance away from the preceding answer box can also lead to errors in following the instructions. They report (p. 17) that 'altering the visual and verbal design of branching instructions has a substantial impact on how well respondents read, comprehend, and act upon the branching instructions'. It follows from this that the

clear location and *visual impact* of instructions are important for successful completion of branching instructions. Most respondents, they acknowledge, do not deliberately ignore branching instructions; they simply are unaware of them. The implications of the findings from Redline *et al*. (2002) are that instructions for branching and skipping should be placed where they are to be used and where they can be seen.

However, with the rise of electronic surveys, problems of instructions, skipping, filtering and branching are reduced immensely, as, depending on the answer given, the computer automatically takes the respondent to the next appropriate place.

We advise judicious and limited use of filtering and branching devices in paper questionnaires. It is particularly important to avoid having participants turning pages forwards and backwards in a questionnaire in order to follow the sequence of questions that have had filters and branches following from them.

24.5 Asking sensitive questions

Sudman and Bradburn (1982) draw attention to the important issue of including sensitive items in a questionnaire. Whilst the anonymity of a questionnaire and, frequently, the lack of face-to-face contact between the researcher and the respondents in a questionnaire might facilitate responses to sensitive material, the issues of sensitivity and threat cannot be avoided, as they might lead to under-reporting (non-disclosure and withholding data) or over-reporting (exaggeration) by participants. Some respondents may be unwilling to disclose sensitive information, particularly if it could harm themselves or others. Why should they share private matters (e.g. about family life and opinions of school managers and colleagues) with a complete stranger (Cooper and Schindler, 2001, p. 341)? Details of age, income, educational background, qualifications and opinions can be regarded as private and/or sensitive matters.

Sudman and Bradburn (1982, pp. 55–6) identify several important considerations in addressing potentially threatening or sensitive issues, for example socially undesirable behaviour (e.g. drug abuse, sexual offences, violent behaviour, criminality, illnesses, employment and unemployment, physical features, sexual activity, behaviour and sexuality, gambling, drinking, family details, political beliefs, social taboos). They suggest that:

- Open rather than closed questions might be more suitable to elicit information about socially undesirable behaviour, particularly about their frequency.
- Long rather than short questions might be more suitable for eliciting information about socially undesirable behaviour, particularly about their frequency.
- Using familiar words might increase the number of reported frequencies of socially undesirable behaviour.
- Using data gathered from informants, where possible, can enhance the likelihood of obtaining reports of threatening behaviour.
- Deliberately loading the question so that overstatements of socially desirable behaviour and understatements of socially undesirable behaviour are reduced might be a useful means of eliciting information.
- With regard to socially undesirable behaviour, it might be advisable first to ask whether the respondent has engaged in that behaviour previously, and then move to asking about his or her current behaviour. By contrast, when asking about socially acceptable behaviour the reverse might be true, i.e. asking about current behaviour before asking about everyday behaviour.
- In order to defuse threat, it might be useful to locate the sensitive topic within a discussion of other more or less sensitive matters, in order to suggest to respondents that this issue might not be too important.
- It is useful to have alternative ways of asking standard questions, for example, sorting cards, or putting questions in sealed envelopes, or repeating questions over time (this has to be handled sensitively, so that respondents do not feel that they are being 'checked'), in order to increase reliability.
- Asking respondents to keep diaries can increase validity and reliability.
- At the end of an interview-based questionnaire, it is useful to ask respondents their views on the sensitivity of the topics that have been discussed.
- It is important to find ways of validating the data.

Sudman and Bradburn suggest (p. 86) that as questions become more threatening and sensitive, it is wise to expect greater bias and unreliability. They draw attention to the fact (p. 208) that several nominal, demographic details might be considered threatening by respondents. This has implications for their location within the questionnaire (discussed below). The issue here is that sensitivity and threat are to be viewed through the eyes of respondents rather than the questionnaire designer; what might appear innocuous to the researcher might be highly sensitive or offensive to participants. We refer readers to Chapter 13 on sensitive educational research.

24.6 Avoiding pitfalls in question writing

Though there are several kinds of questions available, there are some caveats about the framing of questions in a questionnaire:

1. *Don't assume respondent knowledge, opinions or viewpoints*. There is often an assumption that respondents will have the information or have an opinion about the matters in which researchers are interested. This is a dangerous assumption. It is particularly a problem when administering questionnaires to children, who may write anything rather than nothing. This means that the opportunity should be provided for respondents to indicate that they have no opinion, or that they don't know or don't have the answer to a particular question, or that they feel the question does not apply to them. This is frequently a matter in surveys of customer satisfaction in social science, where respondents are asked, for example, to answer a host of questions about the services provided by utility companies (electricity, gas, water) about which they have no strong feelings, and, in fact, they are only interested in whether the service is uninterrupted, reliable, cheap, easy to pay for and that their complaints are solved.
2. *Don't assume that respondents understand difficult terms*. It is essential that, regardless of the type of question asked, the language and the concepts behind the language should be within the grasp of the respondents. Simply because the researcher is interested in, and has a background in, a particular topic is no guarantee that the respondents will be like-minded. The effect of the questionnaire on the respondent has to be considered carefully.
3. *Don't assume respondent interest in, or concern about, the questionnaire*. Just because the researcher is interested in the topic does not mean that the respondent will be at all concerned about the topic, interested in it or concerned about its supposed importance and meaningfulness. The researcher has to make the questionnaire interesting and motivating.
4. *Avoid leading questions*. Avoid questions which are worded (or their response categories presented) in such a way as to suggest to respondents that there is only one acceptable answer, and that other responses might or might not gain approval or disapproval respectively. For example: 'Do you prefer abstract, academic-type courses, or down-to-earth, practical courses that have some pay-off in your day-to-day teaching?' The guidance here is to check the 'loadedness' or possible pejorative overtones of terms or verbs.
5. *Avoid highbrow questions*, even with sophisticated respondents, for example: 'What particular implications of the current positivistic/interpretive debate would you like to see reflected in a course of developmental psychology aimed at a teacher audience?' Where the sample being surveyed is representative of the whole adult population, misunderstandings of what researchers take to be clear, unambiguous language are commonplace. Therefore it is important to use clear and simple language.
6. *Avoid complex questions*, for example: 'Would you prefer a short, non-award bearing course (3, 4 or 5 sessions) with part-day release (e.g. Wednesday afternoons) and one evening per week attendance with financial reimbursement for travel, or a longer, non-award bearing course (6, 7 or 8 sessions) with full-day release, or the whole course designed on part-day release without evening attendance?'
7. *Avoid irritating questions*. Avoid irritating, insulting, embarrassing questions or instructions, for example: 'Have you ever attended an in-service course of any kind during your entire teaching career?' 'If you are over forty, and have never attended an in-service course, put one tick in the box marked NEVER and another in the box marked OLD.'
8. *Avoid complicated instructions*. If your instructions are too difficult to understand at first glance then the respondent could well give up. Golden rule: make the instructions clear, simple and easy to understand.
9. *Avoid negatives and double negatives*. Avoid questions that use negatives and double negatives (Oppenheim, 1992, p. 128), for example: 'How strongly do you feel that no teacher should enrol on the in-service, award-bearing course who has not completed at least two years full-time teaching?' Or: 'Do you feel that without a parent/teacher association or committee teachers are unable to express their views to parents clearly?' In this case, if you feel that a parent/teacher association or committee *is* essential for teachers to express their views, do you vote 'yes' or 'no'? The hesitancy involved in reaching such a decision and the possible required re-reading of the question could cause the respondent simply to leave it blank and move on to the next question. The problem is the double negative: 'without' and 'unable'; it creates confusion.

10 *Avoid too many open-ended questions on self-completion questionnaires*. Because self-completion questionnaires cannot probe respondents to find out just what they mean by particular responses, open-ended questions are problematic. (This caution does not hold in the interview situation, however.) Open-ended questions, moreover, are demanding of most respondents' time and take a lot of time for researcher analysis. Nothing can be more off-putting than the following format:

Use pages 5, 6 and 7 respectively to respond to each of the questions about your attitudes to in-service courses in general and your beliefs about their value in the professional life of the serving teacher.

11 *Avoid extremes in rating scales*, for example, 'never', 'always', 'totally', 'not at all', unless there is a good reason to include them. Most respondents are reluctant to use such extreme categories (Anderson and Arsenault, 1998, p. 174).

12 *Avoid pressuring/biasing by association*, for example: 'Do you agree with your headteacher/principal that boys are more troublesome than girls?' In this case the reference to the headteacher/principal should simply be excised.

13 *The base-rate problem*. Avoid statements with which people generally tend to either disagree or agree, i.e. that have built-in skewedness: the 'base-rate' problem, in which natural biases in the population affect the sample results.

14 *Avoid ambiguous questions*. Avoid ambiguous questions or questions that could be interpreted differently from the way intended. The problem of ambiguity in words is intractable; at best it can be minimized rather than eliminated altogether. The most innocent of questions is replete with ambiguity. Take the following examples:

- Does your child regularly do homework?

What does 'regularly' mean – once a day; once a year; once a term; once a week?

- How many students are there in the school?

What does this mean: on roll, on roll but absent; marked as present but out of school on a field trip; at this precise moment or this week (there being a difference in attendance between a Monday and a Friday), or between the first term of an academic year and the last term of the academic year for secondary school students as some of them will have left school to go into employment and others will be at home revising for examinations or have completed them?

- How many computers do you have in school?

What does this mean: present but broken; including those out of school being repaired; the property of the school or staff's and students' own computers; on average or exactly in school today?

- Have you had a French lesson this week?

What constitutes a 'week': the start of the school week (i.e. from Monday to a Friday), since last Sunday (or Saturday depending on one's religion), or, if the question were put on a Wednesday, since last Wednesday; how representative of all weeks is this week, there being public examinations in the school for some of the week?

It is essential to ensure that questions and their reference are explicit, specific and concrete.

15 *Have discrete categories*. Ensure that categories do not overlap, for example:

- How old are you?

15–20
20–30
30–40
40–50
50–60

Here the categories are not discrete; will an old-looking forty-year-old flatteringly put himself in the 30–40 category, or will an immature twenty-year-old seek the maturity of being put into the 20–30 category? The rule in questionnaire design is to avoid any overlap of categories.

16 *Ask one question at a time*. Ensure that each question only asks about one point. Consider, for example:

- Vocational education is only available to the lower-ability students but it should be open to every student.

This is, in fact, a double question. What does the respondent do who agrees with the first part of the sentence – 'vocational education is only available to the lower-ability students' – but disagrees with the latter part of the sentence, or vice versa? The rule in questionnaire design is to ask only one question at a time.

Though it is impossible to legislate for the respondents' interpretation of wording, the researcher, of course, has to adopt a common-sense approach to this, recognizing the inherent ambiguity but nevertheless still feeling that it is possible to live with this indeterminacy. Piloting can also identify ambiguities and differences of interpretation.

An ideal questionnaire possesses the same properties as a good law, being clear, unambiguous and practicable, reducing potential errors in participants and data analysts, being motivating for participants and ensuring as far as possible that respondents are telling the truth.

17 *Minimize satisficing, acquiescence and social desirability*. Krosnick (1991, 1999) found that the more difficult a question is, the greater is the likelihood of 'satisficing', i.e. choosing the first reasonable response option in a list, rather than working through the list methodically to find the most appropriate or authentic response category. Krosnick and Presser (2010, p. 265) comment that in answering questions, respondents first have to understand and interpret what the question means and is seeking; second, they have to search their memories and minds for relevant information and responses; third, they have to put this all together in coming to a single judgement; and finally they have to convert this judgement into the response. Given these demands, the authors suggest that there is a risk of satisficing, taking shortcuts to give an answer, being less thorough and thoughtful in each of these four stages, giving a satisfactory or what they deem to be an acceptable rather than an accurate answer or even any answer rather than no answer. Krosnick and Presser also comment (p. 271) that the use of a mid-point in a scale may encourage satisficing. Satisficing may also become an issue if the questionnaire is long, as respondent fatigue sets in (p. 292).

Krosnick and Presser (2010, p. 275) note the risk of acquiescence, i.e. where respondents tend to agree with the statement being made, regardless of its content. This may arise for several reasons, for example a wish not to be oppositional or confrontational or to disagree, or a wish to be polite. They note that acquiescence is common in people in lower social positions, with lower intelligence, less formal education, less willingness to think deeply, less concern to present a socially desirable response and where a question is demanding. This places upon researchers the need to ensure that their questions are easy to understand and answer, clear, motivating and neutrally worded.

Respondents may also give an answer in terms of what they think is socially desirable, rather than what they really feel. This leads to bias in answers given, and unreliability. It can be attenuated by careful wording. For example, instead of asking teachers 'how many times have you been absent from school because of stress?', one could ask 'have you ever felt under pressure to come to school even when you knew it was going to be stressful?', and then have a follow-up question: 'has this ever led to you absenting yourself from school?'

The golden rule is to keep questions and questionnaires as short, simple, interesting and easy to complete as possible.

24.7 Sequencing questions

To some extent, the order of questions is a function of the target sample (e.g. how they will react to certain questions), the purposes of the questionnaire (e.g. to gather facts or opinions), the sensitivity of the research (e.g. how personal and potentially disturbing the issues are) and the overall balance of the questionnaire (e.g. where best to place sensitive questions in relation to less threatening questions, and how many of each to include).

The ordering of the questionnaire is important, for early questions may set the tone or the mindset of the respondent to later questions. For example, a questionnaire that makes a respondent irritated or angry early on is unlikely to have managed to enable that respondent's irritation or anger to subside by the end of the questionnaire. As Oppenheim remarks (1992, p. 121), one covert purpose of each question is to ensure that the respondent will continue to cooperate.

Further, a respondent might 'read the signs' in the questionnaire, seeking similarities and resonances between statements so that responses to early statements will affect responses to later statements and vice versa. Whilst multiple items may act as a cross-check, this very process might be irritating for some respondents.

Krosnick and Alwin (1987) report a 'primacy effect' (discussed earlier), i.e. respondents tend to choose items that appear earlier in a list rather than those that appear later in a list. The key principle, perhaps, is to avoid creating a mood-set or a mindset early on in the questionnaire. For this reason it is important to commence the questionnaire with non-threatening questions that respondents can readily answer. After that it might be possible to move towards more personalized questions.

Similarly, the recency effect can bias a response (discussed earlier: respondents remember the last item in a list, rather than the entire list in, for example, a multiple-choice question or a rating scale, particularly in a telephone interview). Hence, for example, a telephone questionnaire should contain short rather than long lists of choices.

Completing a questionnaire can be seen as a learning process in which respondents become more at home with the task as they proceed. Initial questions should

therefore be simple, have high interest value and encourage participation. This will build up the confidence and motivation of the respondent. The middle section of the questionnaire should contain the difficult questions; the last few questions should be of high interest in order to encourage respondents to return the completed schedule.

A common sequence of a questionnaire is:

- Commence with unthreatening factual questions (that, perhaps, will give the researcher some nominal data about the sample).
- Move to closed questions (e.g. dichotomous, multiple choice, rating scales, constant sum questions) about given statements or questions, eliciting responses that require opinions, attitudes, perceptions, views.
- Move to more open-ended questions (or, maybe, intersperse these with more closed questions) that seek responses on opinions, attitudes, perceptions and views, together with reasons for the responses given. These responses and reasons might include sensitive or more personal data.
- Close with potentially sensitive demographic questions, for example, age, qualifications, income.

The move is from objective facts to subjective attitudes and opinions through justifications and to sensitive, personalized data. Clearly the ordering is neither as discrete nor as straightforward as this. For example, an apparently innocuous question about age might be offensive to some respondents; a question about income is unlikely to go down well with somebody who has just become unemployed; and a question about religious belief might be seen as an unwarranted intrusion into private matters. Many questionnaires keep questions about personal details until the end.

The issue here is that the questionnaire designer has to anticipate the sensitivity of the topics in terms of the respondents, and this has a large socio-cultural dimension. What is being argued here is that the *logical* ordering of a questionnaire has to be mediated by its *psychological* ordering. The instrument has to be viewed through the eyes of the respondent as well as the designer.

In considering the sequence of the questionnaire items, then, there are some straightforward guidelines:

- put general, non-threatening questions first;
- make the first questions easy, interesting and able to be answered;
- put important items in the first half of the questionnaire;
- put sensitive or potentially embarrassing questions later in the questionnaire;
- move from factual to abstract questions over the course of the questionnaire;
- put open questions later rather than earlier;
- put demographic and personal questions at the end of the questionnaire.

24.8 Questionnaires containing few verbal items

The discussion so far has assumed that questionnaires are entirely word-based. This might be off-putting for many respondents, particularly children (Smith and Haslett, 2016). In these circumstances a questionnaire might include visual information and ask participants to respond to this (e.g. pictures, cartoons, diagrams), or might include some projective visual techniques (e.g. to draw a picture or diagram, to join two related pictures with a line, to write the words or what someone is saying or thinking in a 'bubble' picture), to tell the story of a sequence of pictures together with personal reactions to it.

The issue here is that in tailoring the format of the questionnaire to the characteristics of the sample, a very wide embrace might be necessary to take in non-word-based techniques. This is not only a matter of *appeal* to respondents, but also, perhaps more significantly, a matter of *accessibility* of the questionnaire to the respondents, i.e. a matter of reliability and validity.

24.9 The layout of the questionnaire

The appearance of the questionnaire is important (e.g. Diaz de Rada, 2005; Dillman *et al*., 2014). It must look easy, attractive and interesting rather than complicated, unclear, forbidding and boring. A compressed layout is uninviting and it clutters everything together; a larger questionnaire with plenty of space for questions and answers is more encouraging to respondents. Verma and Mallick (1999, p. 120) suggest, for paper-based questionnaires, the use of high-quality paper if funding permits.

Layout can be a particular problem in Internet surveys where the screen size is much smaller than the length of a printed page.

Dillman *et al*. (1999, 2014) found that respondents tend to expect less of a form-filling task than is actually required. They expect to read a question, read the response, make a mark and move on to the next question, but in many questionnaires it is more complicated than this. The rule is simple: keep it as uncomplicated as possible.

It is important for respondents to be introduced to the purposes of each section of a questionnaire, so that they can become involved in it and maybe identify with it. If space permits, it is useful to tell the respondent the purposes and focuses of the sections/of the questionnaire, and the reasons for the inclusion of the items. Here Champagne (2014, p. 72) argues for specificity in stating the purposes, avoiding such generalized statements of purpose as 'this questionnaire will help us to improve the quality of our teaching' as this is little more than a statement of the obvious (why else would you be conducting the questionnaire on teaching?).

Clarity of wording and simplicity of design are essential. Clear instructions should guide respondents – 'Put a tick', for example, invites participation, whereas complicated instructions and complex procedures intimidate respondents. Putting ticks in boxes by way of answering a questionnaire is familiar to most respondents, whereas requests to circle pre-coded numbers at the right-hand side of the questionnaire can be a source of confusion and error. This is useful for paper-based questionnaires; with computer-based questionnaires, the use of radio buttons and check boxes renders making a choice very easy.

In some paper-based questionnaires it might also be useful to include an example of how to fill in the questionnaire (e.g. ticking a box, circling a statement), though, clearly, care must be exercised to avoid leading the respondents to answering questions in a particular way by dint of the example provided (e.g. by suggesting what might be a desired answer to the subsequent questions). Verma and Mallick (1999, p. 121) suggest the use of emboldening to draw the respondent's attention to significant features.

Ensure that short, clear instructions accompany each section of the questionnaire. Repeating instructions as often as necessary is good practice in a postal questionnaire. Since everything hinges on respondents knowing exactly what is required of them, clear, unambiguous instructions, boldly and attractively displayed, are essential.

Clarity and presentation also impact on the numbering of the questions. For example a four-page questionnaire might contain sixty questions, broken down into four sections. It might be off-putting to respondents to number each question (1–60) as the list will seem interminably long, whereas to number each section (1–4) makes the questionnaire look manageable. Hence it is useful, in the interests of clarity and logic, to break down the questionnaire into subsections with section headings, and group together similar items or topics. This will also indicate the overall logic and coherence of the questionnaire to the respondents, enabling them to 'find their way' through the questionnaire. It might be useful to preface each subsection with a brief introduction that tells them the purpose of that section. In an Internet questionnaire, numbering the questions may even become redundant, as the computer sets out the questions.

The practice of sectionalizing and sub-lettering questions (e.g. Q9 (a) (b) (c)…) is a useful technique for grouping together questions about a specific issue. It is also a way of making the questionnaire look smaller than it actually is!

The questionnaire designer (particularly in a paper-based questionnaire) must make it clear if respondents are exempted from completing certain questions or sections of the questionnaire (discussed earlier in the section on skips, branches and filters). If so, then it is vital that the sections or questions are numbered so that the respondent knows exactly where to move to next. Here the instruction might be, for example: 'if you have answered "yes" to question 10 please go to question 15, otherwise continue with question 11', or: 'if you are the school principal please answer this section, otherwise proceed to Section 3'.

Arrange the contents of the questionnaire in such a way as to maximize cooperation. For example, include questions that are likely to be of general interest. Make sure that questions which appear early in the format do not suggest to respondents that the enquiry is not intended for them. Intersperse attitude questions throughout the questionnaire to allow respondents to air their views rather than merely describe their behaviour. Such questions relieve boredom and frustration as well as providing valuable information in the process.

Coloured pages and screens can help to clarify the overall structure of the questionnaire and the use of different colours for instructions can assist respondents (this is easily done in an electronic questionnaire). Further, as we noted in Chapter 17, Diaz de Rada (2005) reports that the design, size and colour of the paper used affects response rates. For paper-based questionnaires, small-sized questionnaires were mostly returned by males and those under sixty-four years of age (p. 69), whilst larger-sized questionnaires were mostly returned by females and those over the age of sixty-five (p. 70). He recommends the use of paper size 14.85×21 cm (i.e. a sheet of A4-sized paper folded in half), with white paper and a cover page (p. 73). He reports that paper size has no effect on the quality of the responses.

It is important to include in the questionnaire, perhaps at the beginning, assurances of confidentiality, anonymity and non-traceability, for example by indicating that respondents need not give their name, that

the data will be aggregated, that individuals will not be able to be identified or traced through the use of categories or details of their location etc. (i.e. that it will not be possible to put together a traceable picture of the respondents through the combining of nominal, descriptive data about them). In some cases, however, the questionnaire might ask respondents to put their names so that they can be traced for follow-up interviews in the research (Verma and Mallick, 1999, p. 121); here the guarantee of eventual anonymity and non-traceability will still need to be given (and this applies also to data archiving, where identifying data are removed).

Redline *et al*. (2002) indicate that the placing of the response categories to the immediate right of the text increases the chance of it being answered (the visual *location*), and making the material more salient (e.g. through emboldening and capitalization) can increase the chances of it being addressed (the *visibility* issue). This is particularly important for branching questions and instructions.

They also note that questions placed at the bottom of a page tend to receive more non-response than questions placed further up on the page. Indeed they found that putting instructions at the bottom of the page, particularly if they apply to items on the next page, can easily lead to those instructions being overlooked. It is important, then, to consider what should go at the bottom of the page, perhaps the inclusion of less important items at that point. The authors suggest that questions with branching instructions should not be placed at the bottom of a page. Though Redline *et al*. wrote about paper-based questionnaires, their comments can also apply to Internet-based questionnaires and screen layout.

Finally, a brief note at the very end of the questionnaire can: (a) ask respondents to check that no answer has been inadvertently missed out; (b) solicit an early return of the completed schedule; (c) thank respondents for their participation and cooperation.

24.10 Covering letters/sheets and follow-up letters

The purpose of the covering letter/sheet is to indicate the aim of the research, to convey to respondents its importance, to assure them of confidentiality and to encourage their replies. The covering letter/sheet should:

- provide a title to the research;
- introduce the researcher, giving her/his name, address, organization, contact telephone/fax/email address, together with an invitation to feel free to contact the researcher for further clarification or details;
- indicate the purposes of the research;
- indicate the importance and benefits of the research;
- indicate why the respondent has been selected for receipt of the questionnaire;
- indicate any professional backing, endorsement or sponsorship of, or permission for, the research (e.g. university, professional associations, government departments). The use of a logo can be helpful here;
- set out how to return the questionnaire (e.g. in the accompanying stamped, addressed envelope, in a collection box in a particular institution, to a named person; whether the questionnaire will be collected – and when, where and by whom) (for an Internet questionnaire return might be automatic);
- indicate the address to which to return the questionnaire;
- indicate what to do if questions or uncertainties arise (e.g. a helpline);
- indicate a return-by/complete-by date;
- indicate any incentives for completing the questionnaire;
- provide assurances of confidentiality, anonymity and non-traceability;
- indicate how the results will and will not be disseminated, and to whom;
- thank respondents in advance for their cooperation.

Verma and Mallick (1999, p. 122) suggest that, where possible, it is useful to personalize the letter, avoiding 'Dear colleague', 'Dear Madam/Ms/Sir' etc., and replacing these with exact names.

With these intentions in mind, the following practices are recommended:

- The appeal in the covering letter must be tailored to suit the particular audience. Thus, a survey of teachers might stress the importance of the study to the profession as a whole.
- Neither the use of prestigious signatories, nor appeals to altruism, nor the addition of handwritten postscripts affects response levels to postal questionnaires.
- The name of the sponsor or the organization conducting the survey should appear on the letterhead as well as in the body of the covering letter.
- A direct reference should be made to the confidentiality of respondents' answers and the purposes of any serial numbers and codings should be explained.

- A pre-survey letter advising respondents of the forthcoming questionnaire has been shown to have a substantial effect on response rates (Dillman *et al*., 2014).
- A short covering letter is most effective; no more than one page. An example of a covering letter for teachers and senior staff is set out in Boxes 24.1 and 24.2.

For a further example of a questionnaire, see the accompanying website.

24.11 Piloting the questionnaire

The wording of questionnaires is of paramount importance and pre-testing is crucial to their success (cf. Krosnick and Presser, 2010; Dillman *et al*., 2014; Owen *et al*., 2016). A pilot has several functions, principally to increase the reliability, validity and practicability of the questionnaire (Oppenheim, 1992; Morrison, 1993; Wilson and McLean, 1994, p. 47; Verma and Mallick, 1999, p. 120; Krosnick and Presser, 2010; Dillman *et al*., 2014):

- to check the clarity of the questionnaire items, instructions and layout;
- to gain feedback on the validity of the questionnaire items, the operationalization of the constructs and the purposes of the research;
- to eliminate ambiguities or difficulties in wording;
- to check readability levels for the target audience;
- to gain feedback on the *type* of question and its format (e.g. rating scale, multiple choice, open, closed etc.);
- to gain feedback on response categories for closed questions and multiple-choice items, and for the appropriateness of specific questions or stems of questions;
- to identify omissions and redundant and irrelevant items;
- to gain feedback on leading questions;
- to gain feedback on the attractiveness and appearance of the questionnaire;
- to gain feedback on the layout, sectionalizing, numbering and itemization of the questionnaire;
- to check the time taken to complete the questionnaire;
- to check whether the questionnaire is too long or too short, too easy or too difficult;
- to generate categories from open-ended responses to use as categories for closed response modes (e.g. rating scale items);
- to identify how motivating/non-motivating/sensitive/threatening/intrusive/offensive items might be;
- to identify redundant questions, for example, those questions which consistently gain a total 'yes' or 'no' response, i.e. those questions with little discriminability;
- to identify which items are too easy, too difficult, too complex or too remote from the respondents' experience;

BOX 24.1 EXAMPLE OF A COVERING LETTER

Dear colleague,

IMPROVING SCHOOL EFFECTIVENESS

We are asking you to take part in a project to improve school effectiveness, by completing this short research questionnaire. The project is part of your school development, support management and monitoring of school effectiveness, and the project will facilitate a change management programme that will be tailor-made for the school. This questionnaire is seeking to identify the nature, strengths and weaknesses of different aspects of your school, particularly in respect of those aspects of the school over which the school itself has some control. It would be greatly appreciated if you would be involved in this process by completing the sheets attached, and returning them to me. Please ***be as truthful as possible*** in completing the questionnaire.

You do not need to write your name, and no individuals will be identified or traced from this, i.e. confidentiality and anonymity are assured. If you wish to discuss any aspects of the review or this document please do not hesitate to contact me. I hope that you will feel able to take part in this project.

Thank you.

Signed

Contact details (address, fax, telephone, email)

BOX 24.2 A SECOND EXAMPLE OF A COVERING LETTER

Dear colleague,

PROJECT ON CONDUCTING EDUCATIONAL RESEARCH

I am conducting a small-scale piece of research into issues facing researchers undertaking investigations in education. The topic is very much under-researched in education, and that is why I intend to explore the area.

I am asking you to be involved as you yourself have conducted empirical work as part of a Master's or doctorate degree. No one knows the practical problems facing the educational researcher better than you.

The enclosed questionnaire forms part of my investigation. May I invite you to spend a short time in its completion?

If you are willing to be involved, please complete the questionnaire and return it to XXX by the end of November. You may either place it in the collection box at the General Office at my institution or send it by post (stamped addressed envelope enclosed), or by fax or email attachment.

The questionnaire will take around fifteen minutes to complete. It employs rating scales and asks for your comments and a few personal details. You do not need to write your name, and you will not be able to be identified or traced. When completed, I intend to publish my results in an education journal.

If you wish to discuss any aspects of the study then please do not hesitate to contact me.

I very much hope that you will feel able to participate. May I thank you, in advance, for your valuable cooperation.

Yours sincerely,

Signed

Contact details (address, fax, telephone, email)

- to identify commonly misunderstood or non-completed items (e.g. by studying common patterns of unexpected response and non-response);
- to try out the coding/classification system for data analysis.

In short, as Oppenheim (1992, p. 48) remarks, *everything* about the questionnaire should be piloted; nothing should be excluded, not even the typeface or the quality of the paper (see also Krosnick and Presser, 2010).

The above outline describes a particular kind of pilot: one that does not focus on data, but on matters of coverage and format, gaining feedback from a limited number of respondents and experts on the items set out above.

There is a second type of pilot. This is one which starts with a long list of items and, through statistical analysis and feedback, reduces those items (Kgaile and Morrison, 2006). For example, a researcher may generate an initial list of, for example, 120 items to be included in a questionnaire, and wish to know which items to excise. A pilot is conducted on a sizeable and representative number of respondents (e.g. 50–100) and this generates real data – numerical responses – which can be analysed for:

a *reliability*: those items with low reliability can be removed (Cronbach's alpha for internal consistency: see Chapter 40);
b *collinearity*: if items correlate very strongly with others then a decision can be taken to remove one or more of them, provided, of course, that this does not result in the loss of important areas of the research (i.e. human judgement prevails over statistical analysis);
c *multiple regression*: those items with low betas (see Chapter 42) can be removed, provided, of course, that this does not result in the loss of important areas of the research (i.e. human judgement must prevail over statistical analysis);
d *factor analysis*: to identify clusters of key variables and to identify redundant items (see Chapter 43).

As a result of such analysis, the items for removal can be identified, and this can result in a questionnaire of manageable proportions. It is important to have a good-sized and representative sample here in order to generate reliable data for statistical analysis; too few respondents in this type of pilot may result in important items being excluded from the final questionnaire.

24.12 Practical considerations in questionnaire design

Drawing together the issues discussed so far in questionnaire design, a range of practical implications for designing a questionnaire can be highlighted (cf. Black, 1999; Krosnick and Presser, 2010; Abascal and Diaz de Rada, 2014; Champagne, 2014; Denscombe, 2014; Dillman *et al*., 2014; Hilton, 2017). Sellitz and her associates (1976) have provided a useful guide to researchers in constructing their questionnaires which we summarize in Box 24.3.

Operationalization

- Operationalize the purposes of the questionnaire carefully.
- Ensure that the data acquired will cover the topics and research questions comprehensively and answer the research questions, and that the information asked for is relevant, for example, facts, opinions, behaviour, events, attitudes etc.
- Ensure that every issue has been explored exhaustively; decide on the content and explore it in depth and breadth.
- Use several items to measure a specific attribute, concept or issue.

BOX 24.3 A GUIDE FOR QUESTIONNAIRE CONSTRUCTION

A Decisions about question content

1. Is the question necessary? Just how will it be useful?
2. Are several questions needed on the subject matter of this question?
3. Do respondents have the information necessary to answer the question?
4. Does the question need to be more concrete, specific and closely related to the respondent's personal experience?
5. Is the question content sufficiently general and free from spurious concreteness and specificity?
6. Do the replies express general attitudes and only seem to be as specific as they sound?
7. Is the question content biased or loaded in one direction, without accompanying questions to balance the emphasis?
8. Will the respondents give the information that is asked for?

B Decisions about question wording

1. Can the question be misunderstood? Does it contain difficult or unclear phraseology?
2. Does the question adequately express the alternative with respect to the point?
3. Is the question misleading because of unstated assumptions or unseen implications?
4. Is the wording biased? Is it emotionally loaded or slanted towards a particular kind of answer?
5. Is the question wording likely to be objectionable to the respondent in any way?
6. Would a more personalized wording of the question produce better results?
7. Can the question be better asked in a more direct or a more indirect form?

C Decisions about form of response to the question

1. Can the question best be asked in a form calling for a check answer (or short answer of a word or two, or a number), free answer or check answer with a follow-up answer?
2. If a check answer is used, which is the best type for this question – dichotomous, multiple-choice ('cafeteria' question) or scale?
3. If a checklist is used, does it cover adequately all the significant alternatives without overlapping and in a defensible order? Is it of reasonable length? Is the wording of items impartial and balanced?
4. Is the form of response easy, definite, uniform and adequate for the purpose?

D Decisions about the place of the question in the sequence

1. Is the answer to the question likely to be influenced by the content of preceding questions?
2. Is the question led up to in a natural way? Is it in correct psychological order?
3. Does the question come too early or too late from the point of view of arousing interest and receiving sufficient attention, avoiding resistance, and so on?

Source: Adapted from Sellitz *et al.* (1976)

Respondents

- Consider respondent effort and load: avoid overloading respondents with thinking, recalling, reading, responding; avoid placing too great a burden/demand on respondents in answering the question.
- Consider the reading, writing, listening and thinking abilities of the respondents.
- Consider respondent motivation and ability to answer.
- Consider the willingness of the respondent to answer the questions correctly, accurately and honestly, and whether the respondent will actually have the answer (e.g. to factual questions or to questions which require long-term memory). Remember that respondents may not know the answer or their recall may be faulty.
- Relevance: make sure that the questions included actually apply to the respondents.
- Ensure that the wording is comprehensible to the respondent (use easy words) and judge how the respondent will regard and feel about the question asked.

Ethics

- Address informed consent, right not to take part and to withdraw.
- Address privacy, confidentiality, anonymity, non-traceability.
- Do no harm.
- Address ethical issues in data archiving.
- Consider respondent reactions and effects on respondents.
- Bring beneficence.

Order

- Consider the order of the questions (they are not independent of each other, and the answer to one question may affect the answer to another in the respondent's mind, e.g. primacy, 'carry over' and 'anchoring' effects (Dillman *et al*., 2014, p. 235) (what comes first affects what comes later and respondents use the early questions as a standard against which they compare the later questions)).
- Make the order and organization of the questionnaire easy for the respondent to understand (subheadings in a written survey are useful here).
- Start with a question that is meaningful, interesting and salient to the respondents.
- Make the early questions interesting, able to be answered and easy to answer.
- Group together questions that cover similar topics, to make understanding easy, with subheadings in written surveys, to parallel what would naturally happen in a conversation (as, if respondents see two questions as similar, then they are likely to give similar answers).
- Put important questions in the first half of the questionnaire, and avoid putting them at the end of the survey (later responses may suffer from respondent fatigue, which leads to satisficing).
- Within each topic area, proceed from the general to the specific.
- Put sensitive questions later in the questionnaire in order to avoid creating a mental set in the mind of respondents, but not so late in the questionnaire that boredom and lack of concentration have set in.
- Intersperse sensitive questions with non-sensitive questions.
- If you are using branching questions, ask all the branching questions before asking the follow-up questions.

Question planning

- Ensure that the question actually applies to the respondent.
- Ensure that the question is necessary and relevant for the research purposes and research questions. Remove redundant items ruthlessly.
- Consider what the question is asking for, for example, factual answers; attitudes, perceptions and opinions; behaviours; events; and how to make these clear to the respondent.
- Do not assume that respondents know the answers, or have information to answer the questions, or will always tell the truth (wittingly or not). Include 'don't know', 'not applicable', 'unsure', 'neither agree not disagree' and 'not relevant' categories if appropriate.
- Remember that some factual information is easy (e.g. gender, age) but other data (e.g. attitudes, behaviours and those which rely on memory) may be less accurate.
- Remember that some factual personal questions may be sensitive, so place them at the end of the questionnaire.
- Ensure a balance of questions asking for facts and opinions.

Question type

- Decide on the most suitable and appropriate *type* of question, for example: (a) for *nominal variables*: dichotomous, multiple choice (single choice, restricted number of choices, e.g. three from a longer list, free number of choices); for *ordinal variables*: rating scales, ranking scales; (c) for *interval, ratio and continuous variables*: constant sum,

percentages, marks out of 10, open number (e.g. number of hours of study in a week); (d) for *non-numerical answers*: open questions.
- Frame questions with the data analysis in mind, plan so that the appropriate scales and kinds of data (e.g. nominal, ordinal, interval and ratio) are used.
- Ask more closed than open questions for ease of analysis (particularly with large samples).

Question construction and wording

- Consider the readability levels of the questionnaire and match them to the respondents.
- Use simple, clear language and syntax.
- Avoid jargon; use simple, factual, familiar and non-technical terms.
- Keep the questions (and instructions) simple, clear and short as possible, with as few words as possible, but no fewer.
- Make the wording as concrete, specific, precise, unambiguous and as clear as possible, so that the respondent understands exactly what is being asked for in the questionnaire.
- Avoid making the questions too hard.
- Ask only one thing at a time in a question.
- Use a single, complete, easily structured sentence per item wherever possible.
- Keep statements in the present tense wherever possible.
- Balance brevity with politeness.
- Avoid being offensive.
- Avoid leading questions (those which influence the response and indicate a desired response).
- Try to avoid threatening and embarrassing questions, or write them as neutrally as possible.
- Balance the number of negative questions with the number of positive questions.
- Avoid negatively worded items.
- Avoid double negatives.
- Ensure that the questions are accurate and that the metrics are appropriate (Kosnick and Presser (2010) give the example of a question which measured the height of a horse in feet, whereas it should be in 'hands' (units of four inches)).
- Ensure that the questions use the appropriate scales of measurement and scales (e.g. 1–5, −4 to +4, 'strongly disagree' to 'strongly agree').
- Decide whether to have a mid-point in scale items.
- Note that having a mid-point often leads to greater reliability.
- Note that the absence of a mid-point may force responses into unwelcome choices.
- Use large range scales if subsequent factor analysis is intended or if nuanced responses are required (NB scales higher than seven-point make little difference to the nuancing and may overwhelm respondents).
- Consider the ordering of the scales in each question (e.g. positive to negative; negative to positive, placing a low or high score on the left).
- Avoid double-barrelled questions (asking more than one thing in a single question).
- Take steps to reduce satisficing, acquiescence and social desirability in responses.
- Decide how to avoid falsification of responses (e.g. introduce a checking mechanism into the questionnaire responses to another question on the same topic or issue).

Response categories

- Include sufficient response categories and ensure that they are exhaustive, to fit the choices that participants will really want, i.e. to enable respondents to say what they want to say (which underlines the importance of running a pilot).
- Make response categories discrete (no overlaps), with not too many choices.
- Keep response categories simple and short.
- Consider including a mid-point.
- Clarify to respondents the kinds of responses required in open questions.
- Ensure that the respondent knows how to enter a reply to each question, for example, by underlining, circling, ticking, writing, checking a box.

Length

- Consider the length of the questionnaire; long questionnaires may suffer from respondent fatigue, which leads to satisficing.
- Balance comprehensiveness and exhaustive coverage of issues with the demotivating factor of having respondents complete several pages of a questionnaire.

Layout and instructions

- Make the layout of the questionnaire very clear, unambiguous and attractive.
- Avoid splitting an item over more than one page (or one screen in Internet questionnaires), as the respondent may think that the item from the previous page is finished.
- Avoid putting all the instructions at the start of the questionnaire.
- Keep the instructions close to the questions involved.
- Avoid putting instructions at the foot of a page (or, for electronic surveys, on a different screen from the question to which it applies).

- Provide instructions for introducing, completing and returning (or collection of) the questionnaire (provide a stamped addressed envelope if it is to be a postal questionnaire).

Response rate

- Be satisfied if you receive a 50 per cent response to the questionnaire (a very much lower response rate is probably going to be the case).
- Decide what you will do with missing data and what is the significance of the missing data (that might have implications for the strata of a stratified sample targeted in the questionnaire), and why the questionnaires have not been completed and returned (e.g. were the questions too threatening? Was the questionnaire too long? This might have been signalled in the pilot).
- Consider imputation methods for missing data (see Chapter 17).

Covering letter

- Include a covering letter (or screen, for electronic surveys) with explanation, thanking the potential respondent for anticipated cooperation, indicating the purposes of the research, how anonymity and confidentiality will be addressed, who you are and what position you hold, who will be party to the final report, and your contact details.

Administration

- Consider the medium of the administration and conduct, for example, postal service, email, face-to-face interview, website, telephone, i.e. the visual, oral and aural administration of the survey and who enters the responses (the respondent or the interviewer).
- Decide: whether you (the researcher) will be present when the questionnaire is being completed; whether it is advisable to have an interviewer/researcher present or absent, as the interviewer's/researcher's presence may bias the respondent, raising issues of the respondent's concern for (a) social desirability and (b) acquiescence, and acquiescence is a particular problem in questions which include 'agree', as there is a tendency to agree.
- If the questionnaire is going to be administered by someone other than the researcher, ensure that instructions for administration are provided and that they are clear.

Pre-piloting and piloting

- Be prepared to have a pre-pilot (often with open questions) to generate items for a pilot questionnaire, and then be ready to modify the pilot questionnaire for the final version.
- Pilot the questionnaire, using a group of respondents who are drawn from the possible sample but who will not receive the final, refined version.
- If the pilot includes many items, and the intention is to reduce the number of items through statistical analysis or feedback, then be prepared to have a second round of piloting, after the first pilot has been modified.

A key issue that permeates this lengthy list is for the reader to pay considerable attention to respondents, to see the questionnaire through their eyes and envisage how they will regard it (e.g. from hostility to suspicion to apathy to grudging compliance to welcome; from easy to difficult, from motivating to boring, from straightforward to complex etc.). Address 'brevity, simplicity and concreteness' (Hilton, 2017, p. 30).

24.13 Administering questionnaires

Questionnaires can be administered in several ways, including:

- self-administration
- post
- face-to-face interview (individual and group)
- telephone
- drop-off (see Chapter 17)
- Internet.

Here we discuss only self-administered and postal questionnaires. Chapter 25 covers administration by face-to-face interview and telephone, and administration by the Internet, and we also refer readers to Chapters 17 and 18 on surveys, both paper-based and Internet-based.

The setting in which the questionnaire is completed can also exert an influence on the results. Strange *et al*. (2003, p. 343) found that asking students to complete a questionnaire in silence in a classroom or in a hall set out in an examination style might be very challenging for some; some students did not want to complete a questionnaire 'on their own' and wanted clarification from other students, some wanted a less 'serious' atmosphere to prevail whilst completing the questionnaire, and some (often boys) simply did not complete a questionnaire in conditions that they did not like (p. 344). Researchers must consider how best to achieve reliability by taking into account the setting and preferences of the respondents, and, in the case of schools (p. 345), this includes:

- the timing of the completion;
- the school timetable;
- the space available;
- the layout of the room;
- the size of the school;
- the relationships between the students and the researchers;
- the culture of the school and classrooms;
- the duration of lessons.

Self-administered questionnaires

Self-administration questionnaires are widely used. There are two types here: those that are completed in the presence of the researcher and those that are filled in when the researcher is absent (e.g. at home, in the workplace). We recall the work of Krosnick and Presser (2010) earlier, in indicating the demands made upon respondents in terms of reading, understanding and interest in the question, searching their memories, integrating their information into a judgement and translating their judgement into a response. Self-administration brings all of these four points into sharp relief (Duckworth and Yeager, 2015, p. 240), particularly for some school students or low-achievers. Researchers must decide whether his or her presence is useful or counter-productive.

Self-administered questionnaires in the presence of the researcher

The presence of the researcher may be helpful in enabling any queries or uncertainties to be addressed immediately. Further, it typically ensures a good response rate (e.g. undertaken with teachers at a staff meeting or with students in one or more classes). It can also check that all the questions are completed (the researcher can check these before finally receiving the questionnaire) and filled in correctly (e.g. no rating scale items that have more than one entry per item, and no missed items). It means that the questionnaires are completed rapidly and on one occasion, i.e. it can gather data from many respondents simultaneously.

On the other hand, having the researcher present may be threatening and exert a sense of compulsion, where respondents may feel uncomfortable about completing the questionnaire, and may not wish to complete it or even start it. Respondents may also want extra time to think about and complete the questionnaire, maybe at home, and they are denied the opportunity to do this.

Having the researcher present also places pressure on the researcher to attend at an agreed time and in an agreed place, and this may be time-consuming and require the researcher to travel extensively, thereby extending the time frame for data collection. Travel costs for conducting the research with dispersed samples could also be expensive.

Self-administered questionnaires without the presence of the researcher

The absence of the researcher may be helpful in enabling respondents to complete the questionnaire in private, to devote as much time as they wish to its completion, to be in familiar surroundings and to avoid the potential threat or pressure to participate caused by the researcher's presence. It can be inexpensive to operate, and is more anonymous than having the researcher present. This latter point, in turn, can render the data more (or, indeed, less) honest: it is perhaps harder to tell lies or not to tell the whole truth in the presence of the researcher, and it is also easier to be honest and revealing about sensitive matters without the presence of the researcher.

The down side is that the researcher is not there to address any queries or problems that respondents may have, and respondents may omit items or give up rather than try to contact the researcher. They may wrongly interpret the question and, consequently, answer it inaccurately. They may present an untrue picture to the researcher, for example answering what they would like a situation to be rather than what the actual situation is, or painting a falsely negative or positive picture of the situation or themselves. The absence of the researcher means that the researcher has no control over the environment in which the questionnaire is completed, for example, time of day, noise distractions, presence of others with whom to discuss the questions and responses, seriousness given to the completion of the questionnaire, or even whether it is completed by the intended person.

Postal questionnaires

A postal questionnaire is useful in educational research. Take, for example, the researcher studying the adoption and use made of a new curriculum series of textbooks in secondary schools. An interview survey based upon some sampling of the population of schools would be both expensive and time-consuming. A postal questionnaire, on the other hand, has several distinct advantages. Moreover, given the usual constraints over finance and resources, it might well prove the only viable way of carrying through such an enquiry.

A number of myths about postal questionnaires are not borne out by the evidence (see Krosnick and Presser, 2010; Dillman *et al*., 2014). Response levels to postal surveys are not invariably lower than those obtained by interview procedures; frequently they

equal, and in some cases surpass, those achieved in interviews. Nor does the questionnaire necessarily have to be short in order to obtain a satisfactory response level. With sophisticated respondents, for example, a short questionnaire might appear to trivialize complex issues with which they are familiar. There are several factors in securing a good response rate to a postal questionnaire.

Initial mailing

- use good-quality envelopes, typed and addressed to a named person wherever possible;
- use rapid postage services, with stamped rather than franked envelopes wherever possible;
- enclose a stamped addressed envelope for the respondent's reply;
- in surveys of the general population, Thursday is the best day for mailing out; in surveys of organizations, Monday or Tuesday are recommended;
- avoid at all costs a December survey (questionnaires will be lost in the welter of Christmas postings in the western world).

Follow-up letter

In connection with maximizing response levels, the follow-up letter has been shown to be very effective. The following points should be borne in mind in preparing reminder letters:

- all of the rules that apply to the covering letter apply even more strongly to the follow-up letter;
- the follow-up should re-emphasize the importance of the study and the value of the respondents' participation;
- the use of the second person singular, the conveying of an air of polite disappointment at non-response and some surprise at non-cooperation have been shown to be effective ploys;
- nowhere should the follow-up give the impression that non-response is normal or that numerous non-responses have occurred in the particular study;
- the follow-up letter must be accompanied by a further copy of the questionnaire together with a stamped addressed envelope for its return;
- second and third reminder letters suffer from the law of diminishing returns, so how many follow-ups are recommended and what success rates do they achieve? It is difficult to generalize, but the following points are worth bearing in mind. A well-planned postal survey might obtain a 40 per cent response rate and with the judicious use of reminders, a 70 to 80 per cent response level. A preliminary pilot survey is invaluable in that it can indicate the general level of response to be expected. There is evidence that the use of three reminders can increase the original return by as much as 30 per cent in surveys of the general public. A typical pattern of responses to the three follow-ups is as follows:

Original despatch	40 per cent
First follow-up	+20 per cent
Second follow-up	+10 per cent
Third follow-up	+5 per cent
Total	75 per cent

Bailey (1994, pp. 163–9) shows that follow-ups can be both by mail and by telephone. If a follow-up letter is sent, then this should be around three weeks after the initial mailing. A second follow-up is also advisable (*ibid*.), and this should take place one week after the first follow-up. He reports research (p. 165) that indicates that a second follow-up can elicit up to a 95.6 per cent response rate compared to a 74.8 per cent response with no follow-up. A telephone call *in advance* of the questionnaire can also help in boosting response rates (by up to 8 per cent). More recently, Dillman *et al*. (2014) note that mixed mode questionnaires, particularly with advance notice and follow-up reminders, can be very effective in securing higher response rates.

Incentives

An important factor in maximizing response rates is the use of incentives. It can substantially reduce non-response rates, particularly when the chosen incentives accompany the initial mailing rather than being mailed subsequently as rewards for the return of completed schedules. The explanation of the effectiveness of this appears to lie in the sense of obligation that is created in the recipient. Care is needed in selecting the most appropriate type of incentive. It should clearly be seen as a token rather than a payment for the respondent's efforts and should be as neutral as possible. We refer the reader to discussion of incentives and increasing response rates in Chapter 17.

The preparation of a flow chart can help the researcher to plan the timing and the sequencing of the various parts of a postal survey. One such flow chart, suggested by Hoinville and Jowell (1978), is shown in Figure 24.2. The researcher might wish to add a chronological chart alongside it to help plan the exact timing of the events shown here.

Researchers have to consider, first, whether respondents who complete questionnaires do so accurately, and second, whether those who fail to return their questionnaires would have given the same distribution of

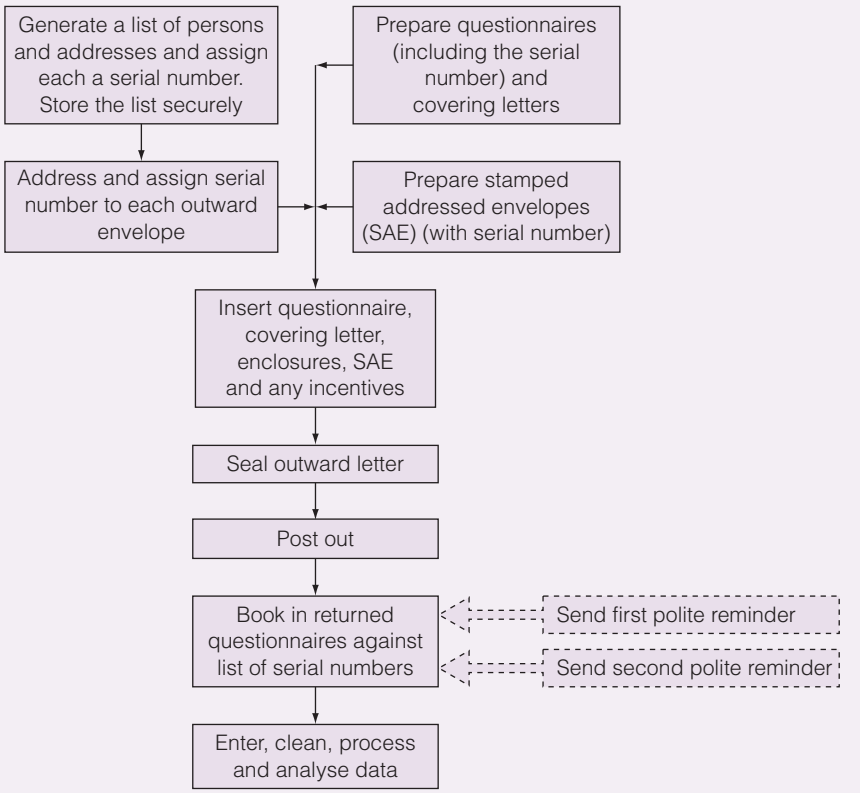


FIGURE 24.2 A flow chart for the planning of a postal questionnaire
Source: Adapted from Hoinville and Jowell (1978)

answers as did the returnees. We discuss the problem of non-response and how to reduce it in Chapter 17.

Further, we devote an entire chapter (Chapter 18) to Internet questionnaires, and we direct readers to this here.

20.14 Processing questionnaire data

Let us assume that researchers have followed the advice we have given about the planning, design and administration of questionnaires and have secured a high response rate to their surveys. Their task is now to reduce the mass of data they have obtained to a form suitable for analysis. Such 'data reduction' generally consists of coding data in preparation for analysis – by hand in the case of small surveys; by computers when the size is greater. First, however, prior to coding, the questionnaires have to be checked. This task is referred to as *editing*.

Editing questionnaires is intended to identify and eliminate errors made by respondents. Moser and Kalton (1977) point to three central tasks in editing:

1. *Completeness*: a check is made that there is an answer to every question. In most surveys, interviewers are required to record an answer to every question (a 'not applicable'/'don't know'/'decline to answer' or 'other' category always being available). Missing answers can sometimes be cross-checked from other sections of the survey. At worst, respondents can be contacted again to supply the missing information. Imputation methods (see Chapter 17) can also be used to compensate for missing data.
2. *Accuracy*: as far as is possible a check is made that all questions are answered accurately. Inaccuracies arise out of carelessness on the part of either interviewers or respondents. Sometimes a deliberate attempt is made to mislead. A tick in the wrong box, a ring round the wrong code, an error in simple arithmetic – all can reduce the validity of the data unless they are picked up in the editing process.
3. *Uniformity*: a check is made that interviewers have interpreted instructions and questions uniformly. Sometimes the failure to give explicit instructions

over the interpretation of respondents' replies leads to interviewers recording the same answer in a variety of answer codes instead of one. A check on uniformity can help eradicate such errors.

The primary task of data reduction is *coding*, that is, assigning a code number to each answer to a survey question. Of course, not all answers to survey questions can be reduced to code numbers (e.g. open-ended questions). Coding can be built into the construction of the questionnaire itself. In this case, we talk of pre-coded answers. Where coding is developed after the questionnaire has been administered and answered by respondents, we refer to post-coded answers. Pre-coding is appropriate for closed questions: male 1, female 2, for example; or single 1, married 2, separated 3, divorced 4. For questions such as those whose answer categories are known in advance, a coding frame is generally developed before the interviewing commences so that it can be printed into the questionnaire itself. It is vital to get coding frames right from the outset – extending them or making alterations at a later point in the study is both expensive and wearisome.

For open-ended, qualitative questions ('Why did you choose this particular in-service course rather than XYZ?'), we refer readers to the discussion of qualitative data analysis in Part 5.

There are several computer packages that will automatically process questionnaire data and return them in useable form (e.g. an Excel file, and SPSS file). At the time of writing some such are SurveyMonkey, FluidSurveys, SphinxSurvey, QuestionPro, SurveyGizmo, Zoho, Typeform, Survey Anyplace. Such packages assist researchers in the design, administration and processing of questionnaires, either for paper-based or for on-screen administration. Responses can be entered rapidly, and data can be examined automatically, producing graphs and tables, as well as a wide range of statistics.

Whilst coding is usually undertaken by the researcher, Sudman and Bradburn (1982, p. 149) also make the case for coding by the respondents themselves, to increase validity. This is particularly valuable in open-ended questionnaire items, though, of course, it does assume not only the willingness of respondents to become involved *post hoc* but also that the researcher can identify and trace the respondents, which, as was indicated earlier, is an ethical matter.

We address data analysis in Part 5.

For considering electronic/Internet questionnaires we refer the reader to Chapter 18.

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Interviews

CHAPTER 25

Interviews are a widely used instrument for data collection. This chapter sets out a range of key issues to be considered in planning, conducting and reporting interviews, including:

- conceptions of the interview
- purposes of the interview
- types of interview
- planning and conducting interviews
- group interviewing
- interviewing children
- interviewing minority and marginalized people
- focus groups
- non-directive, focused, problem-centred and in-depth interviews
- telephone interviewing
- online interviewing
- ethical issues in interviewing

This chapter indicates different kinds of interview, and argues for 'fitness for purpose' to be addressed in deciding which kind of interview and interview questions to employ.

25.1 Introduction

The use of the interview in research marks a move away from seeing human subjects as simply manipulable and data as somehow external to individuals, and towards regarding knowledge as generated between humans, often through conversations (Kvale, 1996, p. 11). Regarding an interview, as Kvale (p. 14) remarks, as an *inter-view*, an interchange of views between two or more people on a topic of mutual interest, sees the centrality of human interaction for knowledge production, and emphasizes the social situatedness of research data.

The interview is a social, interpersonal encounter, not merely a data-collection exercise. Kvale (1996) identifies two different approaches to interviews: the 'miner' who thinks that the interviewee has the information and who is concerned to extract the nuggets of precious material from the interviewee, and the 'traveller' who is concerned to travel with the interviewee as a partner into an unknown country. Whilst the former extracts information, the latter co-constructs knowledge, and this latter more clearly echoes Kvale's view of an interview as an *inter-view*.

As we suggested in Chapter 7, knowledge should be seen as constructed between participants, generating *data* – gifts – rather than *capta* (Laing, 1967, p. 53). As such, the interview is not exclusively either subjective or objective, it is intersubjective (p. 66). Interviews enable participants – interviewers and interviewees – to discuss their interpretations of the world in which they live, and to express how they regard situations from their own point of view. In these senses the interview is not simply concerned with collecting data about life: it is life itself; its human embeddedness is inescapable.

The interview is a flexible tool for data collection, enabling multi-sensory channels to be used: verbal, non-verbal, seen, spoken, heard and, indeed with online interviews, written. The order of the interview may be controlled whilst still giving space for spontaneity, and the interviewer can press not only for complete answers but for responses about complex and deep issues.

Hochschild (2009) notes that the interview can do what surveys cannot, which is to explore issues in depth, to see how and why people frame their ideas in the ways that they do, how and why they make connections between ideas, values, events, opinions, behaviours, etc. They can be used to cast further explanatory insight into survey data, or indeed to set up a survey. In short, the interview is a powerful tool for researchers. On the other hand, interviews are expensive in time, they are open to interviewer bias, they may be inconvenient for respondents, interviewee fatigue may hamper the interview and anonymity may be difficult. We explore these and other issues in this chapter.

An interview is not an ordinary, everyday conversation (Dyer, 1995, pp. 56–8). In contrast to an everyday conversation, it has a specific purpose (to gain evidence or data or information), it is often question-based, with the questions being asked by the interviewer; the interviewer may express ignorance (as may the interviewee), and the responses must be as explicit and as

detailed as possible. The interview is a constructed and usually a specifically planned event rather than a naturally occurring situation, and this renders it different from an everyday conversation; the researcher, therefore, has an obligation to set up, and abide by, the 'rules of the game' in an interview. Indeed Kvale (1996) notes that everyday conversation does not follow a prescripted set of questions, does not occur by appointment and does not have 'respondents' (pp. 30–1).

25.2 Conceptions of the interview

Kitwood (1977) lucidly contrasts three conceptions of an interview. The first conception is that of a potential means of pure information transfer. Here accurate data may be obtained if the interviewer establishes rapport, puts questions in an acceptable manner, and if the respondent is sincere and motivated to answer without lying or giving purely a socially desirable response.

A second conception of the interview is that of a transaction which inevitably has bias which must be recognized and controlled. Here Kitwood explains that each participant – interviewer and interviewee – will define the interview in a particular way. The interview is best understood in terms of a theory of motivation which recognizes a range of non-rational factors governing human behaviour, like emotions, unconscious needs and interpersonal influences. Kitwood points out that both these views of the interview regard inherent features of interpersonal transactions as potential obstructions to research, and, if possible, to be controlled or eliminated.

The third conception of the interview sees it as an encounter necessarily sharing many of the features of everyday life (e.g. Box 25.1). Kitwood suggests that what is required, according to this view, is not a technique for dealing with bias, but a theory of everyday life that takes account of the relevant features of interviews. These may include role-playing, stereotyping, perception and understanding. As Walford (2001, p. 90) remarks, 'interviewers and interviewees co-construct the interview'. As mentioned above, the interview is a social encounter, not simply a site for information exchange or capture, and interviewers should keep this in the forefront of their minds.

One of the strongest advocates of this latter viewpoint is Cicourel (1964), who lists five unavoidable features of the interview situation that would normally be regarded as problematic:

1. There are many factors which inevitably differ from one interview to another, such as mutual trust, social distance and the interviewer's control.
2. The respondent may well feel uneasy and adopt avoidance tactics if the questioning is too deep.
3. Both interviewer and respondent are bound to hold back part of what is in their power to state.
4. Many of the meanings which are clear to one will be relatively opaque to the other, even when the intention is genuine communication.
5. It is impossible, just as in everyday life, to bring every aspect of the encounter within rational control.

The message here is that no matter how hard an interviewer may try to be systematic and objective, the constraints of everyday life will be a part of whatever interpersonal transactions she initiates. Indeed Barker and Johnson (1998, p. 230) argue that the interview is a particular medium for enacting or displaying people's knowledge of cultural forms, as questions, far from being neutral, are couched in the cultural repertoires of all participants, indicating how people make sense of their social world and of each other.

BOX 25.1 ATTRIBUTES OF ETHNOGRAPHERS AS INTERVIEWERS

Trust: There would have to be a relationship between the interviewer and interviewee that transcended the research, promotes a bond of friendship, a feeling of togetherness and joint pursuit of a common mission rising above personal egos.

Curiosity: There would have to be a desire to know, to learn people's views and perceptions of the facts, to hear their stories, discover their feelings. This is the motive force that drives researchers to tackle and overcome the many difficulties involved in setting up and conducting successful interviews.

Naturalness: One endeavours to be unobtrusive in order to see events as they are, untainted by one's presence and actions, so the aim is to secure what is within the minds of interviewees, uncoloured and unaffected by the interviewer.

Source: Adapted from Woods (1986)

25.3 Purposes of the interview

The purposes of the interview are many and varied, for example:

- to understand, evaluate or assess a person, situation or event(s) in some respect;
- to select or promote an employee;
- to effect therapeutic change (e.g. the psychiatric interview);
- to test or develop hypotheses;
- to develop a research instrument such as a survey (as in cognitive interviews (Priede *et al*., 2014));
- to gather data, as in surveys, experimental situations and case studies;
- to sample respondents' opinions.

Although in each of these situations the respective roles of the interviewer and interviewee may vary and the motives for taking part may differ, a common denominator is the transaction that takes place between seeking information on the part of one and supplying information on the part of the other.

As a distinctive research technique, the interview may serve three purposes. First, it may be used as the principal means of gathering information to serve the research objectives, acquiring information on what a person is thinking, knows, likes, values and believes (Tuckman, 1972). Second, it may be used to test hypotheses or to suggest new ones; or to be an explanatory device to help identify variables and relationships. And third, the interview may be used in conjunction with other methods in a research undertaking. In this connection, Kerlinger (1970) suggests that it might be used to follow up unexpected or survey results, for example, or to validate other methods, or to go deeper into the motivations of respondents and their reasons for responding as they do.

Interviews as a research tool range from the formal interview in which set questions are asked and the answers recorded on a standardized schedule through less formal interviews in which the interviewer is free to modify the sequence of questions, change the wording, explain them or add to them, to the completely informal interview where the interviewer may have a number of key issues which she raises in conversational style instead of having a set questionnaire. Beyond this point is located the non-directive interview in which the interviewer takes on a subordinate role.

The research interview has been defined as a conversation between two people which is designed to obtain research data to meet objectives of research which concern 'systematic description, prediction or explanation' (Cannell and Kahn, 1968, p. 527). It involves the gathering of data through direct verbal interaction between individuals and, in this sense, it differs from the questionnaire where the respondent is required to record in some way her responses to set questions (though online written interviews blur these differences).

As the interview has some things in common with the self-administered questionnaire, it is frequently compared with it. Each has advantages over the other in certain respects. The advantages of the questionnaire, for instance, are: it tends to be more reliable because it is anonymous (though this is attenuated in an online interview); it encourages greater honesty; and it is more economical than the interview in terms of time and money (e.g. it can be mailed). Its disadvantages, on the other hand, are: poor response rate; unlike a questionnaire, the interviewer can answer questions concerning both the purpose of the interview and any misunderstandings experienced by the interviewee, as the same questions have different meanings for different people; if only closed items are used, the questionnaire will be subject to the weaknesses discussed in Chapter 24; if only open items are used, respondents may be unwilling to write their answers for one reason or another; questionnaires present problems to people of limited literacy. With the rise of the online interview, some of these distinctions are blurred, and we discuss this below.

We illustrate the relative merits of the interview and the questionnaire in Table 25.1. The direct interaction in the interview is the source of both its advantages and disadvantages as a research technique (Borg, 1963). One advantage, for example, is that it allows for greater depth than is the case with other methods of data collection. A disadvantage, on the other hand, is that it is prone to subjectivity and bias on the part of the interviewer and interviewee.

Oppenheim (1992, pp. 81–2) suggests that interviews have a higher response rate than questionnaires because respondents become more involved and, hence, motivated; they enable more to be said about the research than is usually mentioned in a covering letter to a questionnaire, and they are better than questionnaires for handling more difficult and open-ended questions.

25.4 Types of interview

The number of types of interview is frequently a function of the sources one reads. For example, LeCompte and Preissle (1993) give six types: (i) standardized interviews; (ii) in-depth interviews; (iii) ethnographic

TABLE 25.1 SUMMARY OF RELATIVE MERITS OF INTERVIEW VERSUS QUESTIONNAIRE

Consideration	*Interview*	*Questionnaire*
1 Personal need to collect data	Requires interviewers	Requires a secretary
2 Major expense	Payment to interviewers	Postage and printing
3 Opportunities for response-keying (personalization)	Extensive	Limited
4 Opportunities for asking	Extensive	Limited
5 Opportunities for probing	Possible	Difficult
6 Relative magnitude of data reduction	Great (because of coding)	Mainly limited to rostering
7 Typically, the number of respondents who can be reached	Limited	Extensive
8 Rate of return	Good	Poor
9 Sources of error	Interviewer, instrument, coding, sample	Limited to instrument and sample
10 Overall reliability	Quite limited	Fair
11 Emphasis on writing skill	Limited	Extensive

Source: Tuckman (1972)

interviews; (iv) elite interviews; (v) life history interviews; (vi) focus groups. Bogdan and Biklen (1992) add to this: (vii) semi-structured interviews; (viii) group interviews. Lincoln and Guba (1985) add: (ix) structured interviews; and Oppenheim (1992, p. 65) adds: (x) exploratory interviews. Patton (1980, p. 206) outlines four types: (xi) informal conversational interviews; (xii) interview guide approaches; (xiii) standardized open-ended interviews; (xiv) closed quantitative interviews. Patton sets these out as shown in Table 25.2.

How is the researcher to comprehend the range of these various types? Kvale (1996, pp. 126–7) sets the several forms of interview along a series of continua, arguing that interviews differ in the openness of their purpose, their degree of structure, the extent to which they are exploratory or hypothesis-testing, whether they seek description or interpretation, or whether they are largely cognitive-focused or emotion-focused. A major difference appears to lie in the degree of structure in the interview (cf. Wellington, 2015, p. 141), which itself reflects the purposes of the interview, for example, to generate numbers of responses about a given issue or to indicate unique, alternative feelings about a particular matter. Lincoln and Guba (1985, p. 269) suggest that the structured interview is useful when the researcher is aware of what she does not know and therefore is in a position to frame questions that will supply the knowledge required, whereas the unstructured interview is useful when the researcher is not aware of what she does not know, and therefore relies on the respondents to tell her.

The issue here is of 'fitness for purpose': the more one wishes to gain comparable data – across people, across sites – the more standardized and quantitative one's interview tends to become; the more one wishes to acquire unique, non-standardized, personalized information about how individuals view the world, the more one veers towards qualitative, open-ended, unstructured interviews. This is true not simply of interviews but of their written counterpart – questionnaires. Oppenheim (1992, p. 86) indicates that standardization should refer to *stimulus equivalence*, i.e. that every respondent should *understand* the interview question in the same way, rather than replicating the exact wording, as some respondents might have difficulty with particular questions, or interpret them very differently and perhaps irrelevantly (though he adds that, as soon as the wording of a question is altered, however minimally, it becomes, in effect, a different question).

Oppenheim (1992, p. 65) suggests that *exploratory* interviews are designed to be essentially heuristic and seek to develop hypotheses rather than to collect facts and numbers. He notes that these may cover emotionally loaded topics and, hence, require skill on the part of the interviewer to handle the interview situation, enabling respondents to talk freely and emotionally, with candour, richness, depth, authenticity and honesty in their comments.

Morrison (1993, pp. 34–6) sets out five continua of different ways of conceptualizing interviews. At one end of the first continuum are numbers, statistics, objective facts and quantitative data; at the other end are transcripts of conversations, comments, subjective accounts, essentially word-based qualitative data.

At one end of the second continuum are closed questions, multiple-choice questions where respondents

TABLE 25.2 STRENGTHS AND WEAKNESSES OF DIFFERENT TYPES OF INTERVIEW

Type of interview	*Characteristics*	*Strengths*	*Weaknesses*
1 Informal conversational interview	Questions emerge from the immediate context and are asked in the natural course of things; there is no predetermination of question topics or wording.	Increases the salience and relevance of questions; interviews are built on and emerge from observations; the interview can be matched to individuals and circumstances.	Different information collected from different people with different questions. Less systematic and comprehensive if certain questions don't arise 'naturally'. Data organization and analysis can be quite difficult.
2 Interview guide approach	Topics and issues to be covered are specified in advance, in outline form; interviewer decides sequence and working of questions in the course of the interview.	The outline increases the comprehensiveness of the data and makes data collection somewhat systematic for each respondent. Logical gaps in data can be anticipated and closed. Interviews remain fairly conversational and situational.	Important and salient topics may be inadvertently omitted. Interviewer flexibility in sequencing and wording questions can result in substantially different responses, thus reducing the comparability of responses.
3 Standardized open-ended interviews	The exact wording and sequence of questions are determined in advance. All interviewees are asked the same basic questions in the same order.	Respondents answer the same questions, thus increasing comparability of responses; data are complete for each person on the topics addressed in the interview. Reduces interviewer effects and bias when several interviewers are used. Permits decision makers to see and review the instrumentation used in the evaluation. Facilitates organization and analysis of the data.	Little flexibility in relating the interview to particular individuals and circumstances; standardized wording of questions may constrain and limit naturalness and relevance of questions and answers.
4 Closed quantitative interviews	Questions and response categories are determined in advance. Responses are fixed; respondent chooses from among these fixed responses.	Data analysis is simple; responses can be directly compared and easily aggregated; many short questions can be asked in a short time.	Respondents must fit their experiences and feelings into the researcher's categories; may be perceived as impersonal, irrelevant and mechanistic. Can distort what respondents really mean or experienced by so completely limiting their response choices.

Source: Patton (1980, p. 206)

have to select from a given, predetermined range of responses the particular response which most accurately represents what they wish to have recorded for them; at the other end of the continuum are much more open-ended questions which do not require the selection from a given range of responses, and respondents can answer the questions in their own way and in their own words, i.e. the research is responsive to participants' own frames of reference and response.

At one end of the third continuum is a desire to measure responses, to compare one set of responses with another, to correlate responses, to see how many people said this, how many rated a particular item as such-and-such; at the other end of the continuum is a desire to capture the uniqueness of a particular situation, person or programme – what makes it/them different from others, i.e. to record the quality of a situation or response.

At one end of the fourth continuum is a desire for formality and the precision of numbers and prescribed categories of response where the researcher knows in advance what is being sought; at the other end is a more

responsive, informal intent where what is being sought is more uncertain and indeterminate: we only know what we are looking for when we have found it! The researcher goes into the situation and responds to what emerges.

At one end of the fifth continuum is the attempt to find regularities – of response, opinions etc. – in order to begin to make generalizations from the data, to describe what is happening; at the other end is the attempt to portray and catch uniqueness, the quality of a response, the complexity of a situation, to understand why respondents say what they say, and all of this in their own terms.

One can cluster the sets of poles of the five continua thus:

Quantitative approaches	*Qualitative approaches*
numbers	words
predetermined, given	open-ended, responsive
measuring	capturing uniqueness
short-term, intermittent	long-term, continuous
comparing	capturing particularity
correlating	valuing quality
frequencies	individuality
formality	informality
looking at	looking for
regularities	uniqueness
description	explanation
objective facts	subjective facts
describing	interpreting
looking in from the outside	looking from the inside
structured	unstructured
statistical	ethnographic, illuminative

The left-hand column is much more formal and pre-planned to a high level of detail, whilst the right-hand column is far less formal and the fine detail only emerges once the researcher is *in situ*. Interviews in the left-hand column are front-loaded, that is, they require all the categories and multiple-choice questions to be worked out in advance. This usually requires a pilot to try out the material and refine it. Once the detail of this planning is completed the analysis of the data is relatively straightforward because the categories for analysing the data have been worked out in advance, hence data analysis is rapid.

The right-hand column is much more end-loaded, that is, it is quicker to commence and gather data because the categories do not have to be worked out in advance; they emerge once the data have been collected. However, in order to discover the issues that emerge and to organize the data presentation, the analysis of the data takes considerably longer.

Kvale (1996, p. 30) sets out key characteristics of qualitative research interviews, namely that they should:

- engage, understand and interpret the key feature of the lifeworlds of the participants;
- use natural language to gather and understand qualitative knowledge;
- be able to reveal and explore the nuanced descriptions of the lifeworlds of the participants;
- elicit descriptions of specific situations and actions, rather than generalities;
- adopt a deliberate openness to new data and phenomena, rather than being too pre-structured;
- focus on specific ideas and themes, i.e. have direction, but avoid being too tightly structured;
- accept the ambiguity and contradictions of situations where they occur in participants, if this is a fair reflection of the ambiguous and contradictory situation in which they find themselves;
- accept that the interview may provoke new insights and changes in the participants themselves;
- regard interviews as an interpersonal encounter, with all that this entails;
- be a positive and enriching experience for all participants.

Here we focus on five main kinds of interview that may be used specifically as research tools: (i) the structured interview; (ii) the semi-structured interview; (iii) the unstructured interview; (iv) the non-directive interview; and (v) the focused interview.

In the structured interview the content and procedures are organized in advance, the sequence and wording of the questions are determined by means of a schedule and the interviewer is left little freedom to make modifications. Where some leeway is possible, it, too, is specified in advance. It is characterized by being a closed situation.

In the semi-structured interview, the topics and questions are given, but the questions are open-ended and the wording and sequence may be tailored to each individual interviewee and the responses given, with prompts and probes (discussed below).

The unstructured interview is an open situation, having greater flexibility and freedom. As Kerlinger (1970) notes, although the research purposes govern the questions asked, their content, sequence and wording are entirely in the hands of the interviewer. This does not mean, however, that the unstructured interview is a more casual affair, for in its own way it also has to be carefully planned.

The non-directive interview as a research technique derives from the therapeutic or psychiatric interview.

The principal features of it are the minimal direction or control exhibited by the interviewer and the freedom the respondent has to express her subjective feelings as fully and as spontaneously as she chooses or is able. Moser and Kalton (1977, p. 297) argue that respondents should be encouraged to talk about the subject under investigation (e.g. themselves) and to be free to guide the interview, with few set questions or prefigured frameworks. The interviewer should prompt and probe, pressing for clarity and elucidation, rephrasing and summarizing where necessary and checking for confirmation of his/her understanding of the issue, particularly if the issues are complex or vague.

The need to introduce more interviewer control into the non-directive situation led to the development of the focused interview. The distinctive feature of this type is that it focuses on a respondent's subjective responses to a known situation in which she has been involved and which has been analysed by the interviewer prior to the interview. She is thereby able to use the data from the interview to substantiate or reject previously formulated hypotheses. Here, as Merton and Kendall (1946) explain, 'the interviewer can, when expedient, play a more active role: he can introduce more explicit verbal cues to the stimulus pattern or even *represent* it' (p. 542). It evokes concrete responses by participants.

We examine both the non-directive interview and the focused interview in more detail later in the chapter.

25.5 Planning and conducting interviews

Planning an interview involves sampling, question type, the design of the interview and who is being interviewed. 'How many interviews do I need to conduct?' is a frequent question of novice researchers, asking both about the numbers of people and the number of interviews with each person. Our advice here echoes that of Kvale (1996, p. 101), namely, one conducts interviews with as many people as necessary in order to gain the information sought. There is no simple rule of thumb, as this depends on the purpose of the interview, for example, to make generalizations, to provide in-depth, individual data, to gain a range of responses etc. (see Chapter 12 on sampling for fuller treatment of these matters). The issue here is that the interviewer must ensure that the interviewees selected will be able to furnish the researcher with the information, i.e. that participants actually possess the information.

Kvale (1996, p. 88) sets out several stages in the planning of an interview investigation: thematizing; designing; interviewing; transcribing; analysing; verifying; and reporting. We extend these to a ten-stage sequence to structure our comments here about the planning of interview-based research.

Stage 1: thematizing

The preliminary stage of an interview study is where the purpose of the research is decided. It may begin by outlining the theoretical basis of the study, its broad aims, its practical value and the reasons why the interview approach was chosen. There then follows the translation of the general goals of the research into more detailed and specific objectives and research questions. This is the most important step, for only careful formulation of objectives at this point will eventually produce the right kind of data necessary for satisfactory answers to the research problem.

Stage 2: designing

There follows the preparation of the interview schedule itself. This involves translating the research objectives and research questions into the actual questions that make up the main body of the schedule. This needs to be done in such a way that the questions adequately reflect what the researcher is trying to find out. One can begin this task by writing down the variables to be dealt with in the study (Tuckman, 1972).

Before the actual interview items are prepared, it is important to consider the question format and the response mode. The choice of question format, for instance, depends on a consideration of one or more of the following factors:

- the objectives of the interview;
- the nature of the subject matter;
- whether the interviewer is dealing in facts, opinions or attitudes;
- whether specificity and/or depth is sought;
- the respondent's likely level of understanding;
- the kind of information he or she can be expected to have;
- whether or not the interviewee's thought needs to be structured;
- some assessment of the interviewee's motivational level;
- the extent of the interviewer's own insight into the respondent's situation;
- the kind of relationship the interviewer can expect to develop with the respondent.

From these, the researcher is in a position to decide the kind of interview to adopt, whether to use open and/or closed questions, direct and/or indirect questions, specific and/or non-specific questions, and so on.

Stage 3: construction of schedules

As discussed in detail in Chapter 24, there are several types of question that can be asked: open ended, closed, dichotomous, multiple choice (single response, constrained response – a limited number of choices – and free choice; rank ordering, rating scales, ratio data). We refer the reader to that chapter.

Three main kinds of items are used in the construction of schedules in research interviews (Kerlinger, 1970). First, 'fixed-alternative items' allow the respondent to choose from two or more alternatives, for example, the dichotomous item which offers two alternatives only: 'yes–no' or 'agree–disagree', and sometimes a third alternative such as 'undecided' or 'don't know' is also offered. Kerlinger notes that 'fixed-alternative items' have the advantage of achieving greater uniformity of measurement and therefore greater reliability; of making the respondents answer in a manner fitting the response category; and of being more easily coded. Disadvantages include: their superficiality; the possibility of irritating those respondents who find none of the alternatives suitable; and the possibility of forcing inappropriate responses, either because the alternative chosen conceals ignorance on the part of the respondent or because she may choose an alternative that does not accurately represent the true facts. These weaknesses can be overcome, however, if the items are written with care, mixed with open-ended items and used in conjunction with probes by the interviewer.

Second, 'open-ended items' have been succinctly defined by Kerlinger (1970) as those which provide a frame of reference for respondents' answers, but which put little restraint on the kinds and contents of answers and how they can be expressed, for example, 'What kind of Internet areas do you most search?' Here, other than the subject of the question, which is determined by the nature of the issue under investigation, there are no other restrictions on either the content or the manner of the interviewee's reply.

Open-ended questions have a number of advantages: they are flexible; they allow the interviewer to probe so that she may go into more depth if she chooses, or to clear up any misunderstandings; they enable the interviewer to test the limits of the respondent's knowledge; they encourage cooperation and help to establish rapport; and they allow the interviewer to make a truer assessment of what the respondent really believes. Open-ended situations can also produce unexpected or unanticipated answers which may suggest hitherto unthought-of relationships or hypotheses.

A particular kind of open-ended question is the 'funnel' which, as in questionnaires, starts with a broad question or statement and then narrows down to more specific ones. Kerlinger (1970) quotes an example from the study by Sears *et al*. (1957):

> All babies cry, of course. Some mothers feel that if you pick up a baby every time it cries, you will spoil it. Others think you should never let a baby cry for very long. How do you feel about this? What did you do about it? How about the middle of the night?
>
> (Sears *et al*., 1957, p. 157)

Third, the 'scale' (rating scales) is a set of verbal items to each of which the interviewee responds by indicating degrees of agreement or disagreement (or other anchor statements for response), i.e. the individual's response is located on a scale of fixed alternatives. The use of this technique along with open-ended questions enables scale scores to be checked against data elicited by open-ended questions. It is possible to use one of a number of scales in this context: attitude scales, rank order scales, rating scales, and so on (see also Chapter 24). We touch upon this subject again below.

In devising questions for the interview, Arksey and Knight (1999, pp. 93–5) suggest that attention has to be given to:

- the vocabulary to be used (keeping it simple);
- avoiding prejudicial language;
- avoiding ambiguity and imprecision;
- leading questions (a decision has to be taken whether it is justified to use them);
- avoiding double-barrelled questions (asking more than one point at a time);
- questions which make assumptions (e.g. 'do you go to work in your car?');
- hypothetical or speculative questions;
- sensitive or personal questions (whether to ask or avoid them);
- avoiding assuming that the respondent has the required knowledge/information;
- recall (how easy it will be for the respondent to recall events etc.).

Prompts and probes

The framing of questions for a semi-structured interview can also consider *prompts* and *probes* (Morrison, 1993, p. 66). Prompts enable the interviewer to clarify topics or questions, particularly if the interviewee seems not to have understood, or to have misunderstood, or wishes to ask for clarification or more guidance from the interviewer. The interviewer can, for example, rephrase or repeat the question, or give an example (Denscombe, 2014, p. 193). These have to be

used with caution, as they may falsely assume that the interviewee has not understood the question.

Probes enable the interviewer to ask respondents to extend, elaborate, add to, exemplify, provide detail for, clarify or qualify their response, thereby addressing richness, depth of response, comprehensiveness and honesty that are some of the hallmarks of successful interviewing (see also Patton, 1980, p. 238; Wellington, 2015, p. 147), and they enable the researcher to understand more the thought processes of the interviewee (Priede *et al*., 2014). A probe may be simply the follow-up 'why'/'how' questions or 'can you give me an example of this?'. It could comprise simply repeating the question, repeating the answer in a questioning tone, showing interest and understanding, asking for clarification or an example or further explication, or indeed simply pausing as if to give the interviewee the opportunity to add more.

Aldridge and Levine (2001, p. 119) suggest two types of probe: one in which more detailed factual information is sought, and another in which the respondent is encouraged to elaborate on accounts that they have given or opinions that they hold. Patton (1990) gives three types of probe: detail-oriented probes; elaboration probes; and clarification probes. Beatty and Willis (2007) identify four types of probe (p. 300):

- *anticipated probes*: pre-scripted probes to follow up on an initial question (see also Conrad and Blair's (2009) 'discretionary probes');
- *spontaneous probes*: not pre-scripted, where the interviewer decides on the spur of the moment what to probe, which is not based on a particular response from the interviewee;
- *conditional probes*: pre-scripted probes which are only used if the respondent answers in a particular way or hesitates (see also Conrad and Blair, 2009);
- *emergent probes*: not pre-scripted, where the interviewer decides to probe in response to what the interviewee says, for example, an apparent problem.

Priede *et al*. (2014, p. 560) add three further types:

- *cognitive probes*: focusing on the interviewee's understandings and interpretations of the question, what they drew on and time frames they referred to when answering the question, and how easy or difficult they found the question;
- *confirmatory probes*: 'to check that the information given by the respondent is thus far correct' (p. 560);
- *expansive probes*: seeking further information and details from the interviewees.

Probes can range from the less intrusive (e.g. pausing for the respondent to say more, or making a sound such as 'mmm' to indicate that the interviewer is following closely) to the more intrusive (e.g. repeating a phrase or idea that the respondent said and then following it up with a request for further information, or summarizing ('am I right in thinking that you were saying…', or 'can I just check that I have understood correctly?') and then questioning, or asking for an example or instance, or asking for clarification, or even politely and respectfully challenging, or checking (cf. Aldridge and Levine, 2001, p. 120)).

Fowler (2009, p. 139) offers a cautionary note, suggesting that the more the interviewer prompts and probes, the greater is the chance of bias entering the interview. His argument favours standardized wording, with the possibility of further explanation if respondents are unclear. Wellington (2015, p. 147) advises caution in having too many prompts or probes which ask for depth ('over probing'), as this may provoke resentment or bias.

An interview schedule for a semi-structured interview (i.e. where topics and open-ended questions are written but the exact sequence and wording does not have to be followed with each respondent) might include:

- the topic to be discussed;
- the specific possible questions to be put for each topic;
- the issues within each topic to be discussed, together with possible questions for each issue;
- a series of prompts and probes for each topic, issue and question.

Stage 4: question formats

We now look at the kinds of questions and modes of response in interviewing. First, the matter of question format (cf. Wilson, 1996): how is a question to be phrased or organized? Tuckman (1972) listed four possible formats. For example, questions may take a direct or indirect form. Thus an interviewer could ask a teacher whether she likes teaching: this would be a direct question. Or else she could adopt an indirect approach by asking for the respondent's views on education in general and the ways schools function. From the answers proffered, the interviewer could make inferences about the teacher's opinions concerning her job. Tuckman suggests that by making the purpose of questions less obvious, the indirect approach is more likely to produce frank and open responses.

Second, there are questions which deal with either a general or specific issue. To ask a child what she

thought of the teaching methods of the staff as a whole would be a general or non-specific question. To ask her what she thought of her teacher as a teacher would be a specific question. There is also the sequence of questions, for example, the funnel, in which the movement is from the general and non-specific to the more specific. Tuckman comments that the interviewer must be careful in being too specific too soon, as such direct questions could make the respondent cautious, reticent and avoid an honest answer; rather, coming at an issue more indirectly could produce a more honest response without causing alarm.

Third, a further distinction is made between questions inviting factual answers and those inviting opinions. To ask a person what political party she supports is a factual question. To ask her what she thinks of the current government's foreign policy is an opinion question. Both fact and opinion questions can yield less than the truth, however: the former do not always produce factual answers, nor do the latter necessarily elicit honest opinions. In both instances, inaccuracy and bias may be minimized by careful wording and sequencing of the questions.

There are several ways of categorizing questions, for example (Spradley, 1979; Patton, 1980):

- descriptive questions;
- experience questions;
- behaviour questions;
- knowledge questions;
- construct-forming questions;
- contrast questions (asking respondents to contrast one thing with another);
- feeling questions;
- sensory questions;
- background questions;
- demographic questions.

These concern the *substance* of the question. Kvale (1996, pp. 133–5) adds to these what might be termed *process* questions, i.e. those which:

- introduce a topic or interview;
- follow up on a topic or idea;
- probe for further information or response;
- ask respondents to specify and provide examples;
- directly ask for information;
- indirectly ask for information;
- interpret respondents' replies.

An interviewee may be presented with either a question or a statement. In the case of the latter she will be asked for her response to it in one form or another.

Example question: Do you think homework should be compulsory for all children between nine and twelve years old?

Example statement: Homework should be compulsory for all children between nine and twelve years old.

Agree Disagree Don't know

Stylianou (2008) discusses the 'interview control question'. In experimental and survey designs, variables are often controlled, i.e. held 'constant and unvarying so that one can see the true effects of other variables' after the effects of others have been neutralized (controlled out) (Morrison, 2009, p. 65), i.e. what effects remain after all the other variables have been controlled out. Controlling for the effects of other variables can be undertaken, inter alia, through randomization and random allocation (see Chapter 20), and isolation and control of independent variables other than those in which the researcher is interested (e.g. holding them constant). Controlling for the effects of other variables enables the researcher to see the true effect(s) of a single independent variable in which she or he is interested, i.e. what is left after other variables have been controlled out of the situation. Stylianou (2008) suggests that the same can be done in interviews, i.e. by isolating and controlling out the effects of other variables, the researcher can see the true effect of a particular variable in which she or he is interested, i.e. when the effects of others have been removed. Interview control questions are a form of a probe.

Let us give an example of an interview control question in an imaginary interview concerning a parent who expresses a negative attitude towards mixed ability classes in a primary school:

1. *Interviewer*: Why are you not in favour of mixed ability classes in the school?
2. *Respondent*: The less able students will slow down the more able students in the class, and the teacher will have to work very hard to keep up with the wide range of different abilities in the class.
3. *Interviewer*: But we know that many more capable students slow down anyway, for two reasons: firstly, if they finish work quickly then they are given more work to do, and they want an easy life, and secondly, many of the more able students don't want to stand out as being exceptional in their class, so they slow down. And anyway, the teacher has to work hard, as she has a range of tasks to do as part of her daily work. In fact the teacher can

have a classroom assistant to work with students of different abilities.

4 *Respondent*: But having a classroom assistant still doesn't ensure that all the students get their fair share of the teacher's attention – only the less able and more able children get the extra attention from the classroom assistant.

5 *Interviewer*: But that's the case anyway, as not all the students get the same amount of attention by the teacher, regardless of their abilities, as some students prefer to work quietly on their own without the teacher. Students have to work by themselves anyway, for example, in their mathematics lessons only they can do the work and the teacher cannot do it for them. And, anyway, it's important for students to learn to work by themselves; isn't that a good thing?

6 *Respondent*: But some students aren't good at working by themselves and they may need the teacher's help at critical moments, and having so many mixed abilities will prevent the teacher from being there at critical moments.

7 *Interviewer*: But the teacher will be there to help them at critical moments anyway, that's part of their job, and they are trained to recognize critical moments. Teachers have to be present at critical moments in a student's thinking, prompting them to take the next step in their thinking or learning.

8 *Respondent*: But some students will want to have an easy life, so they won't let the teacher know that they need help or prompting, and they will ask their friend to help them.

9 *Interviewer*: But students do that anyway, as they often help each other; surely that's a good thing, to work collaboratively and help each other, and students learn well from each other.

10 *Respondent*: Look, I just don't want my child to have to work with less able children, and that's all.

In the example, the interviewer is carefully stripping away the possible causes of the parent's attitude: (a) slowing down the more capable students (paras. 2 and 3); (b) students having fair access to adult help (paras. 4 and 5); (c) students having to learn by themselves (paras. 5 and 6); (d) the presence of teachers at critical moments (paras. 6 and 7); (e) students working with friends (paras. 8 and 9). The interviewer, even though being somewhat confrontational (the repeated use of the word 'but'), is raising alternative applications of each of the possible causes, i.e.:

- there are reasons other than the one given here as to why the teacher has to work hard anyway (i.e. not only a matter of having the more and less able students in the same class), and why the more able students may slow down their rate of learning, not only the presence of less able students;
- there are reasons other than the one given here as to why having a classroom assistant will not help to solve the problem of students' access to the teacher's attention (i.e. there are other things that a classroom assistant has to do);
- there are reasons other than the one given here as to why students work by themselves, not only a matter of having or not having the teacher's attention;
- there are reasons other than the one given here as to why the teacher may be present at critical moments;
- there are reasons other than the one given here as to why students work together.

The interviewer is finding that the reasons that the respondent gives for objecting to mixed ability classes operate in other contexts as well, and not solely mixed ability contexts, and so they have to be controlled out: teacher working hard; access to teacher's attention; students working on their own; teacher's non-presence at critical moments; students working collaboratively.

In para. 10, the respondent, having had a range of variables controlled (neutralized) here, becomes exasperated and ends the interview. The researcher might conclude here that the parent is simply prejudiced, other key variables having been removed (controlled) from the reasoning, indeed Stylianou (2008, p. 244) suggests that this kind of probing is useful for studying attitudes and prejudice. Here the interview control question assumes that the variables are dichotomous (e.g. the presence or absence of a variable are the only options); however, within that limitation, the interview control question is useful for identifying possible causal factors in an interviewee's responses.

Stage 5: response modes

Just as there are varied ways of asking questions, so there are several ways in which they may be answered. Here we refer the reader again to Chapter 24 on the several types of question that can be asked and the response modes that accompany them: open-ended, closed, dichotomous, multiple choice (single response, constrained response – a limited number of choices – and free choice); rank ordering, rating scales, ratio data.

As a general rule, the kind of information sought, the means of its acquisition and the kinds of question asked will determine the choice of response mode. The choice of response mode must ensure that the interviewer can be confident that the data will serve her

purposes and analysis of them can be duly prepared. Table 25.3 summarizes the relationship between response mode and type of data.

It is important to bear in mind that more than one question format and more than one response mode can be employed when building up a schedule. The final mixture will depend on the kinds of factors mentioned earlier – the objectives of the research, and so on.

Where an interview schedule is to be used as part of a field survey in which a number of trained interviewers are to be used, it will be necessary to include appropriate instructions for both interviewer and interviewees.

Stage 6: conducting the interview

Setting up and conducting the interview make up the next stage in the procedure. This includes, for example, consideration of the people involved, the location, time and timing of the interview (Mills, 2001), the nature of the interview and the actual conduct of the interview – what happens in it.

Where the interviewer is initiating the research herself, she will clearly select her own respondents; where she is engaged by another agent then she will probably be given a list of people to contact. The interviewer should inform the participant of the nature or purpose of the interview, being honest yet without risking biasing responses (Tuckman, 1972). The interviewer should introduce herself/himself and explain the purposes and conduct of the interview (what happens, and how, and the structure and organization of the interview), how responses may be recorded (and seek permission if this is to happen), and these procedures should be observed throughout (cf. Fowler, 2009, p. 140). The sequence of the question needs careful planning, grouping together similar topics or questions. There are several kinds of question (Kvale, 1996), for example:

- introductory questions (to introduce the topic of that part of the interview, e.g. 'can you tell me about…?');
- follow-up questions (e.g. 'can you tell me a little more about…?'; 'can you give me an example of…?');
- direct questions (e.g. with a 'yes/no' answer);
- indirect questions (to try to obtain the interviewee's real opinion);
- probing questions (to go deeper into a topic);
- specifying questions (e.g. 'what happened next?');
- structuring questions (those that move the interview on to the next topic);
- interpreting questions (to check your understanding, e.g. 'do you mean that…?'; 'Am I right in thinking that…?');
- silence (indicating that you are giving the interviewee the opportunity to expand on a topic or answer).

During the interview the biases and values of the interviewer should not be revealed, and the interviewer must be neutral and avoid being judgemental. The interviewer may have to steer respondents if they are rambling off the point, without being impolite. Aldridge and Levine (2001, p. 119) suggest that factual, personal data should be kept until later in the interview, or at the end, rather than at the beginning of the interview.

The interview, as a social encounter, also requires: clear guidance (however tacit) on when to speak and when to be silent (Mills, 2001, p. 204); tolerance of inattention or lack of clarity; carefully planned greeting and parting; if possible, the topic should put the interviewees at ease (p. 295). In this respect one has to bear in mind that different socio-cultural contexts exert different influences on an interview (not least the linguistic factor in which the researcher may be conducting

TABLE 25.3 THE SELECTION OF RESPONSE MODE

Response mode	*Type of data*	*Chief advantages*	*Chief disadvantages*
Fill-in	Nominal	Less biasing; greater response flexibility	More difficult to score
Scaled	Interval	Easy to score	Time-consuming; can be biasing
Ranking	Ordinal	Easy to score; forces discrimination	Difficult to complete
Checklist or Categorical	Nominal (may be interval when totalled)	Easy to score; easy to respond	Provides less data and fewer options

Source: Tuckman (1972)

the interview in a language that is not his/her first language or the respondent's first language). Miltiades (2008) notes that in some cultures, the influence of culture manifests itself in having not only the presence of other members of an extended family at the interview (p. 281), but those other members actively participating in the interview, giving answers, censoring information (p. 282), interrupting, preventing information from being spoken, and passing comments, i.e. adopting a gate-keeping role (p. 283). As she remarks (p. 282), in some cultures the self is a 'we-self' (rather than an 'I-self') in a collective family, and indeed she notes that the Bengali language has no word for 'private'. Just as the researcher brings his or her own cultural background to the interview, so do the respondents (p. 278), and this might affect the nature, substance and amount of data given, the possible biases towards social desirability of answers (the tendency of respondents to give what they believe will be a socially desirable response, or indeed to self-censor) (p. 283), and indeed in some cultures, the tendency for elders – as authority figures – to give answers rather than the initially targeted interviewees (p. 278). As she remarks (p. 281), in some cultures, the interview becomes a social event.

As the interview, as mentioned already, is a social encounter, a speech act (Austin, 1962), Kvale (1996, p. 125) suggests that an interview follows an unwritten script for interactions, the rules for which only surface if they are transgressed. Hence the interviewer must be at pains to conduct the interview carefully, sensitively and with delicacy (Mills, 2001, p. 286). Kvale (1996, p. 147) adds that, as the researcher is the research instrument, the effective interviewer is not only knowledgeable about the subject matter but is also an expert in interaction and communication. The interviewer will need to establish an appropriate atmosphere such that the participant can feel secure to talk freely. This operates at several levels.

For example, there is the need to address the *cognitive* aspect of the interview, ensuring that the interviewer is sufficiently knowledgeable about the subject matter so that she or he can conduct the interview in an informed manner, and that the interviewee does not feel threatened by lack of knowledge. That this is a particular problem when interviewing children has been documented by Simons (1982) and Lewis (1992), who indicate that children will tend to say anything rather than nothing at all, thereby limiting the possible reliability of the data. The interviewer must also be vigilant to the fact that respondents may not always be what they seem; they may be providing misinformation, telling lies, evading the issue, putting on a front (Walford, 2001, p. 91), settling scores and being malicious. By contrast, it is also an issue in interviewing powerful people (Chapter 13), where the interviewer must be well informed of the subject in question.

Further, the ethical dimension of the interview must be borne in mind, ensuring, for example, informed consent, guarantees of confidentiality, beneficence and non-maleficence (i.e. the interview may be to the advantage of the respondent and will not harm her). Ethics also needs to take account of what is to count as data. For example, it is often after the recorder or video camera has been switched off that the 'gems' of the interview are revealed, or people may wish to say something 'off the record'; the status of this kind of information needs to be clarified before the interview commences. The ethical aspects of interviewing are discussed later in the chapter.

Then there is a need to address the *interpersonal, interactional, communicative and emotional* aspects of the interview. For example, the interviewer and interviewee communicate non-verbally, by facial and bodily expression. Something as slight as a shift in position on a chair might convey whether the researcher is interested, angry, bored, uncomfortable with talking about the issue, agreeing, disagreeing and so on, so the interviewer has to be adept at 'active listening'. We note later in the chapter the challenges and benefits of telephone and online interviewing in depriving participants of visual clues.

Further, the onus is on the interviewer to establish and maintain a good rapport with the interviewee. This concerns being clear, polite, non-threatening, friendly and personable, to the point without being too assertive. It also involves being respectful, for example, some respondents may or may not wish to be called by their first name, family name or title; being dressed too casually may not inspire confidence. Rapport also requires the interviewer to communicate very clearly and positively the purpose, likely duration, nature and conduct and contents of the interview, to give the respondent the opportunity to ask questions, to be sensitive to any emotions in the respondent, to avoid giving any signs of annoyance, criticism or impatience, and to leave the respondent feeling better than, or at least no worse than, she or he felt at the start of the interview. The interviewer must put himself/herself in the shoes of the respondent, and be sensitive to how it must feel to be interviewed. Rapport also means establishing trust, for example, about confidentiality, privacy, anonymity, non-traceability and honesty. It does not mean 'liking' the respondent (Dyer, 1995, p. 62); it means handling the situation sensitively, professionally and ethically.

The interviewer is also responsible for considering the *dynamics* of the situation, for example, how to keep the conversation going, how to motivate participants to discuss their thoughts, feelings and experiences, how to overcome the problems of the likely asymmetries of power in the interview (where the interviewer typically defines the situation, the topic, the conduct, the introduction, the course of the interview and the closing of the interview) (Kvale, 1996, p. 126). As Kvale suggests, the interview is not usually a reciprocal interaction between two equal participants. It is important to keep the interview moving forward, and the interviewer must anticipate and plan for how to achieve this, for example by being clear on what she wishes to find out, asking questions that will elicit the kinds of data sought, giving appropriate verbal and non-verbal feedback to the respondent during the interview. It extends even to considering when the interviewer should keep silent (1996, p. 135).

The 'directiveness' of the interviewer has been scaled by Whyte (1982), where a six-point scale of directiveness and responding was devised (1 = the least directive and 6 = the most directive):

1. Making encouraging noises.
2. Reflecting on remarks made by the informant.
3. Probing on the last remark made by the informant.
4. Probing an idea preceding the last remark by the informant.
5. Probing an idea expressed earlier in the interview.
6. Introducing a new topic.

This is not to say that the interviewer should avoid being too directive or not directive enough; indeed on occasions a confrontational style might yield much more useful data than a non-confrontational style. Further, it may be in the interests of the research if the interview is sometimes quite tightly controlled, as this might facilitate the subsequent analysis of the data. For example, if the subsequent analysis seeks to categorize and classify the responses, then it might be useful for the interviewer to clarify meaning and even suggest classifications during the interview (see Kvale, 1996, p. 130).

Patton (1980, p. 210) suggests that it is important to maintain the interviewee's motivation and interest; hence the interviewer must keep boredom at bay, for example, by keeping to a minimum any demographic and background questions. The issue of the *interpersonal* and *interactional* elements reaches further, for the language of all speakers has to be considered, for example, translating the academic language of the researcher into the everyday, more easy-going and colloquial language of the interviewee, in order to generate rich descriptions and authentic data. Patton goes on to underline the importance of clarity in questioning, and this entails the interviewer finding out what terms the interviewees use about the matter in hand, what terms they use among themselves, and avoiding the use of academic jargon (p. 225). The issue here is not only that the language of the interviewer must be understandable to interviewees but that it must be part of their frame of reference, such that they feel comfortable with it.

Further, the age, gender, race, class, dress, language of the interviewers and interviewees all exert an influence on the interview itself. As Mills (2001, p. 296) remarks, consideration should be given to the potential differentiation of power brought about by the characteristics of the interviewer (e.g. age, profession, social class). Bailey (1994, p. 183) reports that survey interviewers may be female, middle-class white-collar workers, yet those they interview may have none of these characteristics. He reports that having women interviewers elicited a greater percentage of honest responses than having male interviewers (p. 182), that having white interviewers interviewing black respondents yielded different results from having black interviewers interview black respondents (pp. 180–1). He also suggests that interviewers should avoid having specific identity with particular groups or countercultures in their dress (p. 185) as this can bias the interview; rather some unobtrusive clothing should be worn so as to legitimize the role of the interviewer by fitting in with the respondents' expectations of an interviewer's appearance. One can add here that people in power may expect to be interviewed by interviewers in powerful positions and it is more likely that an interview with a powerful person may be granted to a higher-status interviewer (discussed fully in Chapter 13).

The issue extends to consideration of who the interviewer is: an insider or an outsider. Lee (2016) notes that, in some circumstances, for example, members of a particular linguistic, social, cultural or ethnic group, that is, having a natural affinity with the group, may enable greater access and rapport to be obtained than if one is a 'social intruder' (p. 40).

The *sequence* and *framing* of the interview questions will also need to be considered, for example, ensuring that easier and less threatening, non-controversial questions are addressed earlier in the interview in order to put respondents at their ease (see Patton, 1980, pp. 210–11). This might mean that the 'what' questions precede the more searching and difficult 'how' and 'why' questions (though, as Patton reminds us (p. 211), knowledge questions – 'what'-type questions – can be threatening). The interviewer's

questions should be straightforward and brief, even though the responses need not be (Kvale, 1996, p. 132). The interviewer will also need to consider the *kinds* of questions to be put to interviewees, discussed earlier.

There are several problems in the actual conduct of an interview that can be anticipated and, possibly, prevented, ensuring that the interview proceeds comfortably, for example:

- avoiding interruptions from outside (e.g. telephone calls, people knocking on the door);
- minimizing distractions;
- minimizing the risk of 'stage fright' in interviewees and interviewers;
- avoiding asking embarrassing or awkward questions;
- jumping from one topic to another;
- giving advice or opinions (rather than active listening);
- summarizing too early or closing off an interview too soon;
- being too superficial;
- handling sensitive matters (e.g. legal, personal, emotional matters).

Arksey and Knight (1999, p. 53) suggest that the interviewer should:

- appear to be interested;
- keep to the interview schedule in a structured interview;
- avoid giving signs of approval or disapproval of responses received;
- be prepared to repeat questions at the respondent's request;
- be prepared to move on to another question without irritation, if the respondent indicates unwillingness or inability to answer the question;
- ensure that he/she (the interviewer) understands a response, checking if necessary (e.g. 'am I right in thinking that you mean …?');
- if a response is inadequate, but the interviewer feels that the respondent may have more to say, thank the respondent and add 'and could you please tell me …';
- give the respondent time to answer (i.e. avoid answering the question for the respondent).

Gadd (2004, p. 397) reports the importance of how the interviewer responds to the interviewee, as an unsupportive, unsympathetic or negative response (even if not intended) could discourage a respondent from proceeding.

In preparing for the interview, the interviewer must become familiar with the topic in hand, consider the structure and sequence of the interview, plan clear questions that can be put in as few words as possible, plan how to respond (both to the people *qua* people and to what they say), plan how to steer the interview and keep it on track and plan how to balance the different areas of the interview, for example, factual, opinions, feelings, values, background/contextual matters.

There is also the issue of how to record the interview. For example, an audio recorder might be unobtrusive but might constrain the respondent; a video recorder might yield more accurate data but might be even more constraining, with its connotation of surveillance. It might be less threatening not to have any recording, in which case the reliability of the data might rely on the memory of the interviewer (though, as Gadd (2004, p. 384) remarks, memory is *motivated* in nature, and may be subject to selective recall). An alternative might be to have the interviewer make notes *during* the interview, but this could be highly off-putting for some respondents. The issue here is that there is a trade-off between the need to catch as much data as possible and yet to avoid having so threatening an environment that it impedes the potential of the interview situation.

The 'ideal' interview, then, meets several 'quality criteria' (Kvale, 1996, p. 145):

- the extent of spontaneous, rich, specific and relevant answers from the interviewee;
- the shorter the interviewer's questions and the longer the subject's answers, the better;
- the degree to which the interviewer follows up and clarifies the meanings of the relevant aspects of the answers;
- the ideal interview is to a large extent interpreted throughout the interview;
- the interviewer attempts to verify his or her interpretations of the subject's answers in the course of the interview;
- the interview is 'self-communicating' – it is a story contained in itself that hardly requires much extra descriptions and explanations.

People may refuse to be interviewed (Bailey, 1994, pp. 186–7; Cooper and Schindler, 2001, p. 301); they might:

- not give a reason for refusing;
- be hostile to what they see as intrusion;
- hold anti-authoritarian feelings;
- feel that it is a waste of time;
- speak a foreign language;

- take an instant dislike to the interviewer;
- say that they are too busy;
- feel embarrassed or ignorant;
- dislike or feel uncomfortable with the topic under review;
- be afraid of the consequences of participating;
- feel inadequate or that they do not know the right answer.

The onus is on the interviewer to try to overcome these factors, whilst recognizing, of course, that they may be legitimate, in which case no further attempt can be made to conduct the interview. It is important for the interviewer to render the interview as a positive, pleasant and beneficial experience, and to convince the participant of their own worth and the importance of the topic (cf. Solberg (2014) discussing interviewing children). If there is a significant difference between the interviewer and the respondent (e.g. gender, age, ethnicity, race, social status, colour, class), it might be advisable to have another interviewer try to conduct the interview (Morrison, 2013a).

So far the assumption has been that there is only one interviewer present at the interview. There is an argument for having more than one interviewer present, so that one can transcribe or observe features that might be overlooked by the other interviewer whilst the other is engaging the respondent (and these roles have to be signalled clearly to the respondent at the interview), and also to share the interviewing. Joint interviews can provide two versions of the interview – a cross-check – and one can complement the other with additional points, leading to a more complete and reliable record. It also enables one interviewer to observe non-verbal features such as the power and status differentials and social dynamics, and, if there is more than one respondent present at the interview, the relationships between the respondents, for example, how they support, influence, complement, agree and disagree with each other, or indeed contradict each other, the power plays at work, and so on.

On the other hand, having more than one interviewer present is not without its difficulties. For example, the roles of the two interviewers may be unclear to the respondents (and it is the job of the interviewers to make this clear), or it may be intimidating to have more than one interviewer present. Researchers will need to weigh carefully the strengths and weaknesses of having more than one interviewer present, and what their roles will be.

We give readers a list of guidelines for conduct during the interview in Box 25.2.

BOX 25.2 GUIDELINES FOR THE CONDUCT OF INTERVIEWS

- Interviews are an interpersonal matter, a social event.
- Avoid saying 'I want to know …'; the interviewee is doing you a favour, not being interrogated.
- How to follow up on questions/answers.
- How to keep people on track and how to keep the interview moving forward.
- How to show respect.
- How to divide your attention as interviewer and to share out the interviewees' responses – giving them all a chance to speak in a group interview.
- Do you ask everyone in a group interview to give a response to a question?
- If there is more than one interviewer, what are the roles of the 'silent' interviewer, and do the interviewees know the roles of the interviewers?
- Who is looking at whom.
- If you need to look at your watch then maybe comment on this publicly.
- Try not to refer to your interview schedule; if you need to refer to it then comment on this publicly (e.g. 'let me just check that I have covered the points that I wanted').
- Avoid using your pen as a threatening weapon, pointing it at the interviewee.
- Consider your non-verbal communication, eye contact, signs of anxiety, showing respect.
- Give people time to think – don't interrupt yourself if there is silence.
- How to pass over from one interviewer to another and from one interviewee to another if there is more than one interviewer or interviewee.
- How to give feedback and acceptance to the interviewees.
- Should you write responses down – what messages does this give?
- Put yourself in the shoes of the interviewee.

continued

continued

- What are the effects of losing eye contact or of maintaining it for too long?
- Think of your body posture – not too laid-back and not too menacing.
- How to interpret and handle silence.
- Avoid looking away from the respondent if possible.
- Avoid interrupting the respondent.
- Avoid judging the respondent or his/her response.
- The interviewer should summarize and crystallize issues and build on them – that is a way of showing respect.
- How to give signs of acceptance of what people are saying, and how to avoid being judgemental.
- Take care of timing – not too long to be boring.
- Give interviewees the final chance to add any comments, and thank them at the end.
- Plan how to hand over the questions to the next interviewer.
- How to arrange the chairs and tables – do you have tables (they may be a barrier or a protection)?
- Identify who controls the data, and when the control of the data passes from the interviewee to the interviewer.
- What to do with 'off the record' data.
- Take time to 'manage' the interview and keep interviewees aware of what is happening and where it is going.
- Vary the volume/tone of your voice.
- Avoid giving your own view or opinion; be neutral.
- Who is working harder – the interviewer or the interviewee?
- Who is saying more – the interviewer or the interviewee?
- If there is more than one interviewer, how to avoid one interviewer undermining another.
- Think of prompts and probes.
- How to respond to people who say little.
- Consider the social (and physical) distance between the interviewer and interviewee(s).
- Consider the layout of the furniture – circle/oval/straight line or what?
- Have a clear introduction which makes it plain how the interview will be conducted and how the interviewees can respond (e.g. turn-taking).
- Make sure you summarize and crystallize every so often.
- How to handle interviewees who know more about the topic than you do.
- Do you have males interviewing females and vice versa (think of age/gender/race etc. of interviewers and interviewees)?
- Give some feedback to respondents every so often.
- What is the interview doing that cannot be done in a questionnaire?
- If there are status differentials then don't try to alter them in the space of an interview.
- Plan what to do if the interviewee 'turns the tables' and tries to be the interviewer.
- Plan what to do with aggressive or angry interviewees.
- Plan what to do if powerful interviewees don't answer your questions (maybe you need to admit that you haven't understood very well, and ask for clarification, i.e. that it is your fault).
- Be very prepared, so that you don't need to look at your schedule.
- Know your subject matter well.
- If people speak fast then try to slow down everything.
- As an interviewer, you have the responsibility for making sure the interview runs well.

Denscombe (2014, pp. 192–4) provides helpful advice for interviews, noting that a 'good' interviewer must be: attentive and sensitive to the interviewees and their feelings; able to tolerate silences; non-judgemental; effective in the use of prompts, probes and checks; and an effective facilitator in group interviews/focus groups.

Interviewers have to be sensitive to their own effect on the interview. For example, they may fail to secure full cooperation or keep to procedures, they may establish an inappropriate environment (physical, cognitive, emotional, interpersonal), they may be exerting undue influence or pressure on the respondent, or they may be

selective in recording the data; we consider the issue of reliability in Chapter 14.

It is important for the interviewer to explain to the respondent the purpose, scope, nature and conduct of the interview, the use to be made of the data, ethical issues, the likely duration of the interview, i.e. to explain fully the 'rules of the game' so that the interviewee is left in no doubt as to what will happen during and after the interview. The interviewer must introduce herself/himself properly and fully to the respondent (maybe even providing identification). The interviewer has to set the scene appropriately, for example, to say that there are no right and wrong answers, that some of the topics may be deep but that they are not designed to be a test, to invite questions and interruptions, and to clear permission for recording. During the interview it is important, also, for the interviewee to speak more than the interviewer, for the interviewer to listen attentively and to be seen by the respondent to be listening attentively, and for the interviewer to be seen to be at ease with the interview.

What is being suggested here is that the interview, as a social encounter, as a series of speech acts, has to take account of, and plan for, the whole range of other possibly non-cognitive factors that form part of everyday conduct.

Stage 7: transcribing

This is a crucial step in interviewing, for there is the potential for massive data loss, distortion and the reduction of complexity. It has been suggested throughout this chapter that the interview is a social encounter, not merely a data-collection exercise; the problem with much transcription is that it becomes solely a record of data rather than a record of a social encounter. As Powney and Watts (1987) remark:

> Talk is dynamic – a quality it loses as soon as it is collected in any way. It is somewhat … like catching rain in a bucket for later display. What you end up with is water, which is only a little like rain.
>
> (Powney and Watts, 1987, p. 16)

Indeed this problem might have begun at the data-collection stage, for example an audio recording is selective, it filters out important contextual factors, neglecting the visual and non-verbal aspects of the interview (Mishler, 1986) (see also the comments below on online and telephone interviewing). Moreover, frequently non-verbal communication gives important information additional to the verbal communication. Morrison (1993, p. 63) recounts the incident of an autocratic headteacher extolling the virtues of collegiality and democratic decision making whilst shaking her head vigorously from side to side and pressing the flat of her hand in a downwards motion away from herself as if to silence discussion and dismiss any other views. To replace audio recording with video recording might make for richer data and catch non-verbal communication, but this is time-consuming to analyse.

Transcriptions inevitably lose data from the original encounter as it represents the translation from one set of rule systems (oral and interpersonal) to another remote rule system (written language). As Kvale (1996, p. 166) suggests, the prefix *trans* indicates a change of state or form; transcription is selective transformation. Therefore it is unrealistic to pretend that the data on transcripts are anything but *already interpreted* data. As Kvale (p. 167) remarks, the transcript can become an opaque screen between the researcher and the original live interview situation.

Because of these difficulties, there can be no single 'correct' transcription; rather the issue becomes whether, how and how much a transcription is useful. Transcriptions are decontextualized, abstracted from time and space, from the dynamics of the situation, from the live form, and from the social, interactive, dynamic and fluid dimensions of their source; they are frozen.

The words in transcripts are not necessarily as solid as they were in the social setting of the interview. Mishler (1991, p. 260) suggests that data and the relationship between meaning and language are contextually situated; they are unstable, changing and capable of endless reinterpretation.

We are not arguing against transcriptions, rather we are cautioning the researcher against believing that they catch everything that happened in the interview. The researcher might need to ensure that different *kinds* of data are recorded in the transcript of the recording, for example:

- what was being said;
- the tone of voice of the speaker(s) (e.g. harsh, kindly, encouraging);
- the inflection of the voice (e.g. rising or falling, a question or a statement, a cadence or a pause, a summarizing or exploratory tone, opening or closing a line of enquiry);
- emphases placed by the speaker;
- pauses (short to long) and silences (short to long);
- interruptions;
- the mood of the speaker(s) (e.g. excited, angry, resigned, bored, uncomfortable, enthusiastic, committed, happy, grudging);
- the speed of the talk (fast to slow, hurried or unhurried, hesitant to confident);

- how many people were speaking simultaneously;
- whether a speaker was speaking continuously or in short phrases;
- who is speaking to whom;
- indecipherable speech;
- any other events that were taking place at the same time that the researcher can recall.

If the transcript is of a video recording, then this enables the researcher to comment on all of the non-verbal communication that was taking place in addition to the words spoken. The issue here is that it is often inadequate to transcribe only spoken words; other data are important. Of course, as soon as other data are noted, this becomes a matter of interpretation (what is a long pause, what is a short pause, was the respondent happy or was it just a 'front', what gave rise to such-and-such a question or response, why did the speaker suddenly burst into tears?). As Kvale (1996, p. 183) notes, interviewees' statements are not simply collected by the interviewer, they are, in reality, co-authored.

Stage 8: analysing

Once data from the interview have been collected, the next stage involves analysing them, for example, by some form of coding, scoring or content analysis. In qualitative data the analysis is almost inevitably interpretive, hence the data analysis is less a completely accurate representation (as in the numerical, positivist tradition) and more a reflexive, reactive interaction between the researcher and the decontextualized data that are already interpretations of a social encounter.

The researcher has to consider whether to focus on those items which the participant mentions or reiterates the most, or whether to deem important those items that arise when the participant wanders from the point or changes the subject, or – in the case of two respondents – whether they actually mean the same even if they are using the same words to describe similar experiences or the same experience (Gadd, 2004, p. 385). At issue here is the unavoidable integration of analysis and interpretation.

The great tension in data analysis is between maintaining a sense of the holism of the interview and the tendency for analysis to atomize and fragment the data, to separate them into constituent elements, thereby losing the synergy of the whole, and in interviews often the whole is greater than the sum of the parts. There are several stages in analysis, for example:

- generating natural units of meaning;
- classifying, categorizing and ordering these units of meaning;
- structuring narratives to describe the interview contents;
- interpreting the interview data.

These are comparatively generalized stages. Miles and Huberman (1994) suggest several tactics for generating meaning from transcribed and interview data:

- counting frequencies of occurrence (of ideas, themes, pieces of data, words);
- noting patterns and themes (Gestalts), which may stem from repeated themes and causes or explanations or constructs;
- seeing plausibility: trying to make sense of data, using informed intuition to reach a conclusion;
- clustering: setting items into categories, types, behaviours and classifications;
- making metaphors: using figurative and connotative language rather than literal and denotative language, bringing data to life, thereby reducing data, making patterns, decentring the data and connecting data with theory;
- splitting variables to elaborate, differentiate and 'unpack' ideas, i.e. to move away from the drive towards integration and the blurring of data;
- subsuming particulars into the general (akin to Glaser's (1978) notion of 'constant comparison', discussed in Chapter 37 of this volume): a move towards clarifying key concepts;
- factoring: bringing a large number of variables under a smaller number of (frequently) unobserved hypothetical variables;
- identifying and noting relations between variables;
- finding intervening variables: looking for other variables that appear to be 'getting in the way' of accounting for what one would expect to be strong relationships between variables;
- building a logical chain of evidence: noting causality and making inferences;
- making conceptual/theoretical coherence: moving from metaphors to constructs to theories to explain the phenomena.

This progression, though perhaps positivist in its tone, is a useful way of moving from the specific to the general in data analysis. Running through the suggestions from Miles and Huberman (1994) is the importance that they attach to coding, partially as a way of reducing what is typically data overload in qualitative data. They suggest that analysis through coding can be performed both within-site and cross-site, enabling causal chains, networks and matrices to be established, all of these addressing what they see as the

major issue of reducing data overload through careful data display.

Coding has been defined by Kerlinger (1970) as the translation of question responses and respondent information to specific categories for the purpose of analysis. As we have seen, many questions are pre-coded, that is, each response can be immediately and directly converted into a score in an objective way. Rating scales and checklists are examples of pre-coded questions. Coding is the ascription of a category label to a piece of data, with the category label either decided in advance or in response to the data that have been collected.

We discuss coding more fully in Chapter 34, and we refer the reader to that discussion, including the advantages and disadvantages of coding.

Content analysis involves reading and judgement; Brenner *et al.* (1985) set out several sequential steps in undertaking a content analysis of open-ended data:

1. Briefing (understanding the problem and its context in detail).
2. Sampling (of people, including the types of sample sought; see Chapter 12).
3. Associating (with other work that has been done).
4. Hypothesis development.
5. Hypothesis testing.
6. Immersion (in the data collected, to pick up all the clues).
7. Categorizing (in which the categories and their labels must: (a) reflect the purpose of the research; (b) be exhaustive; (c) be mutually exclusive).
8. Incubation (e.g. reflecting on data and developing interpretations and meanings).
9. Synthesis (involving a review of the rationale for coding and an identification of the emerging patterns and themes).
10. Culling (condensing, excising and even reinterpreting the data so that they can be written up intelligibly).
11. Interpretation (making meaning of the data).
12. Writing (including: giving clear guidance on the incidence of occurrence; proving an indication of direction and intentionality of feelings; being aware of what is not said as well as what it said – silences; indicating salience to the readers and respondents).
13. Rethinking.

This process, Brenner *et al*. suggest (1985, p. 144), requires researchers to:

- understand the research brief thoroughly;
- evaluate the relevance of the sample for the research project;
- associate their own experiences with the problem, looking for clues from the past;
- develop testable hypotheses as the basis for the content analysis (the authors name this the 'Concept Book');
- test the hypotheses throughout the interviewing and analysis process;
- stay immersed in the data throughout the study;
- categorize the data in the Concept Book, creating labels and codes;
- incubate the data before writing up;
- synthesize the data in the Concept Book, looking for key concepts;
- cull the data – being selective is important because it is impossible to report everything that happened;
- interpret the data, identifying its meaning and implication;
- write up the report;
- rethink and rewrite: have the research objectives been met?

We discuss content analysis fully in Chapter 34.

Hycner (1985) sets out procedures that can be followed when analysing interview data phenomenologically. We saw in Chapters 1 and 15 that the phenomenologist advocates the study of direct experience taken at face value and sees behaviour as determined by the phenomena of experience rather than by external, objective and physically described reality. In addressing this, Hycner's summary guidelines are as follows:

1. *Transcription*: transcribe the interview recording, noting not only the literal statements but also non-verbal and paralinguistic communication.
2. *Bracketing and phenomenological reduction*: enter the world of the unique individual being interviewed, suspending as far as possible the researcher's own meaning and interpretations.
3. *Listening to the interview for a sense of the whole*: listen to the entire recording several times and read the transcription a number of times in order to provide a context for the emergence of specific units of meaning and themes later on.
4. *Delineating units of general meaning*: thoroughly scrutinize both verbal and non-verbal gestures to elicit the participant's meaning, to crystallize and condense what the respondent has said, using, as far as possible, the interviewee's own words.
5. *Delineating units of meaning relevant to the research question*: once the units of general meaning have been noted, reduce them to units of meaning relevant to the research question.

6 *Training independent judges to verify the units of relevant meaning*: have other researchers carry out the above procedures.
7 *Eliminating redundancies*: check the lists of relevant meaning and eliminate those clearly redundant to others previously listed.
8 *Clustering units of relevant meaning*: determine if any of the units of relevant meaning naturally cluster together; whether there seems to be some common theme or essence that unites several discrete units of relevant meaning.
9 *Determining themes from clusters of meaning*: examine all the clusters of meaning to determine if there is/are one (or more) central theme(s) which expresses the essence of these clusters.
10 *Writing a summary of each individual interview*: go back to the interview transcription and write up a summary of it, incorporating the themes elicited from the data.
11 *Return to the participant with the summary and themes, conducting a second interview*: check to see whether the essence of the first interview has been accurately and fully captured.
12 *Modifying themes and summary*: with the new data from the second interview, look at all the data as a whole and modify or add themes as necessary.
13 *Identifying general and unique themes for all the interviews*: look for the themes common to most or all of the interviews as well as the individual variations: (a) noting if there are themes common to all or most of the interviews; then (b) noting when there are themes unique to a single interview or a minority of the interviews.
14 *Contextualization of themes*: place these themes back within the overall contexts or horizons from which these themes emerged.
15 *Composite summary*: write up a composite summary of all the interviews which accurately capture the essence of the phenomenon being investigated, as experienced by the participants, noting, where relevant, individual differences.

We also refer readers to the chapters on analysing qualitative data in Chapters 32 to 37.

Stage 9: verifying

Chapter 14 discusses at length the issues of reliability, validity and generalizability of the data from interviews, and so these issues are not repeated here. Kvale (1996, p. 237) notes that validation must take place at all stages of the interview-based investigation, set out above. For example: (a) the theoretical foundation of the research must be rigorous and there must be a logical link between such theory and the research questions; (b) all aspects of the research design must be sound and rigorous; (c) the data must be accurate, reliable and valid (with consistency and reliability checks undertaken); (d) the translation of the data from an oral to a written medium must demonstrate fidelity to the key features of the interview situation; (e) data analysis must demonstrate fidelity to the data; (f) validation procedures should be in place and used; (g) the reporting should be fair and seen to be fair by readers.

Here there is no single canon of validity; rather fitness for purpose within an ethically defensible framework should be adopted, giving rise to different kinds of validity for different kinds of interview-based research (e.g. structured to unstructured, qualitative to quantitative, nomothetic to idiographic, generalizable to unique, descriptive to explanatory, positivist to ethnographic, pre-ordinate to responsive).

Stage 10: reporting

The nature of the reporting will be decided to some extent by the nature of the interviewing and the audience. For example, a standardized, structured interview may yield numerical data which may be reported succinctly in tables and graphs, whilst a qualitative, word-based, open-ended interview will yield word-based accounts that take up considerably more space.

Kvale (1996, pp. 263–6) suggests several elements of a report: (i) an introduction that includes the main themes and contents; (ii) an outline of the methodology and methods (from designing to interviewing, transcription and analysis); (iii) the results (the data analysis, interpretation and verification); (iv) a discussion (including possible explanations for the findings).

Figures and tables appear in a typical quantitative report; if the interview is more faithfully represented in words rather than numbers then this presents the researcher with the issue of how to present particular quotations. Here Kvale (p. 266) suggests that direct quotations should: (a) illuminate and relate to the general text whilst maintaining a balance with the main text; (b) be contextualized and be accompanied by a commentary and interpretation; (c) be particularly clear, useful and the 'best' of the data (the 'gems'!); (d) include an indication of how they have been edited; and (e) be incorporated into a natural written style of the report.

For sample interview data, see the accompanying website.

25.6 Group interviewing

One technique of interviewing is that of group interviewing. Watts and Ebbutt (1987) and Leshem (2012), for example, have considered the advantages and disadvantages of group interviewing as a means of collecting data in educational research. The group interview can be cost-efficient, time-efficient, generate a wider range of responses than in individual interviews. Bogdan and Biklen (1992, p. 100) add that group interviews can be useful for gaining an insight into what might be pursued in subsequent individual interviews. There are practical and organizational advantages, too. Pre-arranged groups can be used by teachers with minimum disruption. Group interviews are often more time-saving than individual interviews. The group interview can also bring together people with varied opinions, or as representatives of different collectivities.

Arksey and Knight (1999, p. 76) suggest that having more than one interviewee present can provide two versions of events – a cross-check – and one can complement the other with additional points, leading to a more complete and reliable record. It is also possible to detect how the participants support, influence, complement, agree and disagree with each other, and the relationships between them. On the other hand, one respondent may dominate the interview (p. 76). Further, Arksey and Knight suggest that antagonisms may be stirred up at the interview, individuals may be reticent in front of others, particularly if they are colleagues or if the matter is sensitive. They also suggest that a 'public line' may be offered instead of a more honest, personal response, and indeed participants may collude in withholding information.

Watts and Ebbutt (1987) note that group interviews may be of little use in allowing personal matters to emerge, or where the researcher has to aim a series of follow-up questions at one specific member of the group. As they explain, the dynamics at work among the members of the group may deny access to various personal types of data. Group interviews may produce 'group think', discouraging individuals who hold a different view from speaking out in front of the other group members; they may be 'apprehensive, self-conscious and stressed' (Leshem, 2012, p. 3). Leshem reports that 'the group interview is an intensive social encounter that weaves a complex web of communication styles that may convey ambiguous messages' as interviewers and interviewees bring their own personalities, cultures, values, beliefs and backgrounds to the situation, which 'inevitably affect the way they perform in a constrained reality' (2012, p. 6). As Scheurich (1997) remarks, both interviewers and interviewees bring their own 'conscious and unconscious baggage' to an interview (p. 73).

Further, Lewis (1992) comments on the problem of coding the responses of group interviews. For further guidance on this topic and the procedures involved, we refer the reader to Simons (1982), Arksey and Knight (1999) and Part 5 on qualitative data analysis.

Several issues have to be addressed in the conduct of a group interview, for example:

1. How to divide your attention as interviewer and to share out the interviewees' responses, giving them all a chance to speak in a group interview?
2. Do you ask everyone in a group interview to give a response to a question?
3. How to handle people who are too quiet, too noisy, who monopolize the conversation, who argue and disagree with each other?
4. What happens if people become angry with you or with each other?
5. How to make people be quiet/stop talking whilst being polite?
6. How to handle differences in how talkative people are?
7. How to arrange turn-taking (if appropriate)?
8. Do you ask named individuals questions?
9. How can you have individuals answer without forcing them?
10. How to handle a range of very different responses to the same question?
11. Why have you brought together the particular people in the group?
12. Do you want people to answer in a particular sequence?
13. How to handle different ages, ethnicities, genders, status positions etc. within the group of interviewees?
14. What to do if the more experienced or senior people always answer first in a group interview?
15. As an interviewer, how to be vigilant to pick up on people who are trying to speak?

When conducting group interviews the unit of analysis is the view of the whole group and not the individual member; a collective group response is being sought, even if there are individual differences or a range of responses within the group. This ensures that no individual is either unnecessarily marginalized or subject to blame or being ostracized for holding a different view.

Group interviews are also very useful when interviewing children, and it is to this that we now turn.

25.7 Interviewing children

Children have been regarded as 'the best sources of information about themselves' (Docherty and Sandelowski, 1999, p. 177), but it is important for the interviewer to be able to enter their world and childhood culture and to see the situation through their eyes (p. 177). It is important to understand the world of children through their own eyes rather than the lens of the adult. Children differ from adults in cognitive and linguistic development, attention and concentration span, ability to recall, life experiences, what they consider to be important, status and power (Arksey and Knight, 1999, p. 116). All of these have a bearing on the interview.

Jansen (2015) notes that some children may relish an interview and are eager to participate as it takes them seriously and values their views, experiences and stories, which they do not normally have the opportunity to express in their daily lives, and that the interview is non-evaluative and non-judgemental in nature. Indeed Morrison (2013a) reports a study in which the children left the interview encounter feeling very positive about themselves and the interview, even though it had been an unfamiliar experience.

The interview accords children agency and competency which is sometimes denied in their lives (Jansen, 2015, p. 34). Children are seen as a respected resource, not a problem, and their knowledge is seen by all parties as essential for the research (p. 37). Indeed Solberg (2014) notes, not only with children but more widely, that the interview must respect the point that interviewees have ideas, opinions, information and experiences that the researcher wishes to understand (which also requires the interview to clarify to participants the purposes of the interview) (p. 244).

The interview, as we have mentioned before, is a non-naturally occurring social encounter, a series of speech acts; this is acutely important to keep in mind when interviewing children, as the task is to engage children in a safe context, not to interrogate them or pump them for information. Here the ethical issue of *primum non nocere* – 'first, do no harm' – is paramount (Jansen, 2015). An interview is likely to be an unfamiliar situation for many children, and Morrison (2013a) comments that the task of the interviewer, in a reversal of Blumer's (1969) dictum of 'making the familiar strange', is to 'make the strange familiar'. Morrison (2013a) notes several deliberate actions that the interviewer can take to address this, including, in a situation in which group interviewing with children is conducted in school time and on the school premises:

- have a familiar location (a school classroom with complete privacy and no risk of interruption);
- have the interviewers seen around the school before the interviews;
- ensure a good fit between the culture of the interview and the culture and ethos of the school;
- ensure informed consent;
- guarantee anonymity, privacy, confidentiality and non-traceability;
- put the children at their ease at the start of the interview;
- make the interviews serious yet very good natured, easy, enjoyable and positive;
- create a relaxed, friendly and, at times, humorous atmosphere;
- indicate how important the children are in the research;
- take care with clothing, to respect the children rather than to frighten or over-awe them;
- if the lead interviewer is much older than the children, have a much younger research assistant who deliberately dresses down to be more akin to the children;
- have the interviewers use question-and-answer techniques that the children are well used to in class, i.e. the children have expectations of the adults, and the adults deliberately try to fit those expectations to some extent;
- use the language, genre and register of the children wherever possible;
- take care with question structure, sequence and wording, making them easy to understand, clear, concrete and specific, with one-word answers at first, moving to open-ended answers later in the interview;
- take great care with proxemics, personal space and non-verbal communication, and scrupulously avoid intrusion into personal space;
- give positive feedback on, and thanks for, comments received;
- include voting on various items by the group;
- be acutely alert to hesitancies, non-verbal communication and silences, the emotional and social dimensions of the interview and respond appropriately.

Arksey and Knight (1999, pp. 116–18) indicate that it is important to establish trust with children (cf. Solberg's (2014) comments that the interview is a 'co-production' between children and interviewer, even in a guided interview (p. 244)), to put the child at ease quickly and to help him/her to feel confident, to avoid over-reacting (e.g. if the child is distracted). Putting children at ease extends to the choice of venue and time

(Morrison, 2013a; Leeson, 2014). Conducting an interview in the family home or in school sends signals to children; a child in an abusive family might find school an easier venue than the family home, and a child who has problems at school might prefer a neutral venue. Sometimes the school environment might exert a very positive influence on how seriously the child takes an interview on, say, a school-related matter (Morrison, 2013a), although, of course, it may elicit a perceived socially or institutionally desirable response.

Researchers must also make the interview non-threatening and enjoyable, and avoid making the children feel that they have to explain themselves (Leeson, 2014). Solberg (2104) comments that adults must indicate to the children that they are not like 'ordinary' adults (e.g. in the sense of exerting power over the children) but rather that they (adults) lack the knowledge that the children possess and that they (adults) can learn from the children's experiences, opinions and ideas (p. 244).

Interviewers must use straightforward language and child's language (Danby *et al*., 2011) and ask questions that are appropriate for the age of the child (e.g. keep to the 'here and now' and avoid using 'why', 'when' and 'how' questions with very young children (e.g. below five years old), and ensure that children can understand abstract questions (e.g. with older children)). Krähenbühl and Blades (2006) note the need for questions to be very clear, uncomplicated, concrete, specific and straightforward (see also Wilson and Powell, 2001). The researcher must be mindful, too, that some children will choose to respond orally, whilst others may respond in non-verbal communication (Solberg, 2014, p. 246).

In short, interviewers with children must guide the interviews in such a way as to 'invite rather than overrule the informants' (Solberg, 2014, p. 234). All of these require the researcher to give children time to think, and to combine methods and activities in an interview, for example, drawing, playing, writing, speaking, playing a game, using pictures, doing an enjoyable task (Houssart and Evens, 2011), using newspapers, toys or photographs (cf. Leeson, 2014).

Group interviewing can be useful with children, as well as being economical on researcher time, and it encourages interaction between the group rather than simply a response to an adult's question. Group interviews of children might also be less intimidating for them than individual interviews (Greig and Taylor, 1999, p. 132). Eder and Fingerson (2003, p. 34) suggest that a power and status dynamic is heavily implicated in interviewing children; they have little in comparison to the adult. Indeed, Thorne (1994) uses the term 'kids' rather than 'children', as the former is the term used by the children themselves, whereas 'children', she argues, is a term used exclusively by adults, denoting subordinacy (cf. Eder and Fingerson, 2003, p. 34). Mayall (1999) suggests regarding children as a 'minority group', in that they lack power and control over their own lives. If this is the case, then it is important to take steps to ensure that children are given a voice and an interview setting in which they feel comfortable (cf. Maguire, 2005). Group interviewing is such a setting, taking place in as close to a natural surrounding as possible (Greig and Taylor, 1999, p. 131); indeed Eder and Fingerson (2003, p. 45) report the successful use of a high-status child as the interviewer with a group of children.

Group interviewing with children enables them to reach a consensus or, indeed, to challenge each other and participate in a way that may not happen in a one-to-one, adult–child interview, using language that the children themselves use (Houssart and Evens, 2011, p. 65). For example, Lewis (1992) found that ten-year-olds' understanding of severe learning difficulties was enhanced in group interview situations, the children challenging and extending each other's ideas and introducing new ideas into the discussion. Further, interviewing a group of children together can equalize more the power differentials between interviewer and children (Houssart and Evens, 2011).

Having the interview as part of a routine, everyday activity can also help to make it less unnatural, as can making the interview more like a game (e.g. by using props such as toys and pictures). For example, it could be part of a 'show and tell' or 'circle time' session, or part of group discussion time. The issue here is to try to make the interview as informal as possible. Of course, sometimes it may be more useful to formalize the interview, so that children have a sense of how important the situation is, and they can respond to this positively. It can be respectful to have an informal or, indeed, a formal interview; the former maybe for younger children and the latter for older children (cf. Morrison, 2013a).

Whilst group interviews may be useful with many children, individual interviews with children are also valuable. For example, Eder and Fingerson (2003, pp. 43–4) report the value of individual interviews with adolescents, particularly about sensitive matters, for example, relationships, family, body issues, sexuality, love. Indeed they report examples where individual interviews yielded different results from group interviews with the same people about the same topics, and where the individuals valued greatly the opportunity for a one-to-one conversation.

Interviews with children should try to employ open-ended questions, to avoid a single-answer type of

response (Greig and Taylor, 1999; Wright and Powell, 2006; Powell, 2007), as answers to open-ended questions are usually more accurate than answers to closed questions (Wright and Powell, 2006, p. 317), being respondent-driven and respondent-focused, and they can take greater account of children with limited linguistic or cognitive abilities. Closed questions can lead to response bias in that children may provide answers without thinking (p. 317). Waterman *et al*. (2001) report that children gave clear answers to 'yes/no' closed questions even when such question types were deliberately given in respect of unanswerable questions (i.e. where insufficient information had been given for the question to be answered), in other words the format of the question artificially skewed the response. Clearly, however, specific questions may be needed to elicit specific (e.g. factual) details (Wright and Powell, 2006, p. 320).

Another strategy for interviewing children is to use a projection technique. Here, instead of asking direct questions, the interviewer can show a picture or set of pictures, and then ask the children for their responses to it/them (cf. Greig and Taylor, 1999, pp. 132–3; Dalli and Stephenson, 2010; Hurworth, 2012; Leeson, 2014). For example, a child may first comment on the people's colour in the pictures, followed by their gender, suggesting that colour may be more important in their mind than their gender. This avoids a direct question and may reduce the possibility of a biased answer – where the respondent may be looking for cues as to how to respond. Other projection techniques include the use of dolls or puppets, vignettes, drawings by the interviewees, photographs of a particular scene which the respondents have to comment upon (e.g. what is happening? What should be done here?) (cf. Hurworth, 2012), and the 'guess who' technique (i.e. which people might fit a particular description) (Wragg, 2002, p. 157; see also Dalli and Stephenson, 2010; Hurworth, 2012).

Simons (1982), Lewis (1992), Bailey (1994, pp. 447–9), Breakwell (2000, pp. 245–6), Breakwell *et al*. (2006, pp. 245–6), Brenner (2006), Danby *et al*. (2011), Hurworth (2012) and Leeson (2014) chart some challenges in interviewing children, for example, how to:

- overcome children being easily distracted (e.g. some interviewers provide toys or pictures, or children may be fascinated by something as simple as the researcher's pen, or there may be a passing vehicle outside, and these distract the children);
- avoid the researcher being seen as an authority figure (e.g. a teacher, a parent or an adult in a powerful position);
- understand what children mean and what they say (particularly with very young children);
- gather a lot of information in a short time, children's attention span being limited;
- have children reveal what they really think and feel rather than what they think the researcher wants to hear;
- avoid the interview being seen by the child as a test;
- keep the interview relevant;
- overcome young children's unwillingness to contradict an adult or assert themselves, or, in the case of adolescents, deliberately being oppositional in their views;
- interview inarticulate, hesitant and nervous children;
- get the children's teacher away from the children;
- respond to the child who says something then immediately wishes she hadn't said it;
- elicit genuine responses from children rather than simply responses to the interview situation;
- get beyond the institutional, headteacher's or 'expected' response;
- avoid receiving a socially desirable response;
- ensure that the child is giving a true opinion;
- keep children to the point;
- avoid children being too extreme or destructive of each other's views;
- pitch language at the appropriate level;
- overcome the children taking a question too literally (particularly young children), hence avoid metaphors, similes or analogies;
- enable the children to see a situation through other people's eyes;
- avoid the interview being an arduous bore;
- overcome children's poor memories;
- avoid children being too focused on particular features or situations;
- avoid the situation where the child will say 'yes' to anything addressed (an 'acquiescence bias'), for example, by avoiding 'yes/no' questions in favour of open-ended questions;
- overcome the situation of the child saying anything in order to please, or rather than feel they do not have 'the answer';
- overcome the proclivity of some children to say that they 'don't know' (for a variety of reasons, e.g. they are not interested, they genuinely don't know, they don't understand the question, they think that the interviewer might expect them not to know, they are unwilling to disclose what they do know, they are too shy to speak, they cannot explain themselves very well), or simply to shrug their shoulders and remain silent;

- overcome the problem that some children dominate the conversation in a group interview;
- avoid the problem of children feeling very exposed in front of their friends in a group interview;
- avoid children feeling uncomfortable or threatened (addressed, perhaps, by placing children with their friends);
- avoid children telling lies.

Clearly these problems are not exclusive to children; they apply equally well to adult group interviews. Group interviews require skilful chairing and attention to the physical layout of the room so that everyone can see everyone else and has personal space (which varies in different cultures). Group size is also an issue; too few and it can put pressure on individuals, too large and the group fragments and loses focus. Lewis (1992) indicates that a group of around six or seven is an optimum size, though it can be smaller for younger children. The duration of an interview may not be for longer than around fifteen minutes, and it is important to ensure that distractions are kept to a minimum. Simple language to the point and without ambiguity (e.g. avoiding metaphors) is also important. It is crucial to keep in mind that an interview is a social encounter, and children may be very sensitive to the social dynamics and social context of the interview (Morison *et al*., 2000), and not only its cognitive element (Maguire (2005, p. 4) suggests that 'children have good social radar').

Children will be sensitive to the gender and ethnicity of the interviewer, in fact to many features of the interviewer ('interviewer effects'; Denscombe, 2014, p. 191); the very fact that the interviewer is an adult will affect the interview (Morison *et al*., 2000, p. 113), as power inequalities inhere in adult–child relationships (Solberg, 2014). For further information, we refer the reader to Wilson and Powell (2001) and Mukerji and Albon (2010).

25.8 Interviewing minority and marginalized people

Not all the methods of interviewing set out so far will apply to interviewing marginalized people, i.e. those who are 'on the edge' of society (Barron, 1999), for example, economically, socially or politically, or who are 'invisible': stigmatized groups, the unemployed, the elderly, refugees, asylum seekers, travellers, those with special needs, those with a low status in society, those with limited linguistic, cognitive or intellectual abilities, those whose first language is a minority language, the disabled, the chronically ill, children with cerebral palsy, victims of crime, the oppressed, the subordinate, and so on. Parker and Lynn (2002, p. 13) argue that much educational research has ignored marginalized groups by not addressing their concerns or including them as areas of research, or it regards them as minorities who do not merit research, or, if research is conducted with/on them, it uses culturally inappropriate methods of investigation (p. 13). Similarly, Kelly (2007, p. 22) argues that researchers can no longer 'exclude learning disabled children' from research on the grounds that they pose challenges to conventional research methods.

In interviewing marginalized groups, the interviewer will need to consider greater use of informal, open-ended interviews (which follow the train of thought and response of the respondent and which use age-appropriate and context-appropriate language) rather than highly structured interviews (Swain *et al*., 1998, p. 26). The authors recommend the use of narrative, qualitative and in-depth interviews (discussed later), enabling self-disclosure (both by the interviewer and the interviewee) (Swain *et al*., 1998, p. 26), wherein participants 'tell their stories' in their own words (Barron, 1999, p. 38) and recount their subjective experiences and feelings. This gives them a 'voice', where otherwise they would either not be heard or listened to (see also Swain *et al*., 1998; Leeson, 2014; Solberg, 2014). This accords with the emancipatory potential or intent of research that was set out as a key feature of critical educational research in Chapter 3 (see also Barron, 1999, pp. 40, 44–6). Barron suggests that it is important for the interviewee to feel safe, secure, supported, close to the interviewer, and to know that he/she has the undivided attention of the interviewer (1999, p. 41) and a non-judgmental and non-evaluative stance by the interviewer, with built-in opportunities for respondent validation and clearance (interviewees may wish, upon reflection, to withdraw comments initially made at interview).

The researcher must take care not to exploit what are likely to be (perceived) asymmetries of power in the interview (where the researcher may be regarded as having more power than the respondent) (Swain *et al*., 1998, p. 26). Indeed Swain *et al*. (1998, p. 25) remind researchers that the interviews and research on marginalized groups should bring benefit to them, for example, they do not continue to be exploited or marginalized: the issue of reciprocity.

An interview is a communicative encounter, and, for some marginalized groups (e.g. the physically disabled or those with communication difficulties), this is precisely the challenge to be faced by researchers: how to communicate with those who cannot communicate

easily or at all (e.g. those who cannot speak, elective mutes, the deaf, children with degrees of autism, children with Down's syndrome or attention deficit disorder). Here Kelly (2007, pp. 25–6) notes the use of communication cards, pictorial cards (e.g. which indicate feelings), drawing frames, picture books, toys, puppets, photographs of familiar people or places, respecting and working with – and in – the language used by the participant and keeping within their frame of reference, considering the greater use of 'yes/no' questions than open questions (e.g. for students who cannot speak but who can point). She also advocates the use of projection techniques (p. 28) such as 'three wishes', asking participants to draw a picture to represent the matter in hand and the use of 'feeling cards' (cards with pictures of feelings).

Kelly (2007, p. 24) comments that gaining access to marginalized groups may be difficult, and that in the case of those with disabilities, it is likely to be necessary to gain access through gatekeepers, for example, parents, social workers, health team members, carers and suchlike, and indeed to have them present during the interview or to speak on their behalf (e.g. to protect the respondents' rights directly or to act as proxies/advocates). This is important, as Kelly (p. 23) reports the dangers of 'suggestibility' of participants in interviews, which leads to her comment that skilful and sensitive questioning are essential, drawing on participants' actual experiences and taking care not to project the researcher's own interests. Bourne-Day and Treweek (2008) also indicate that issues of privacy and identity may be significant in researching marginalized groups.

In conducting interviews, Kelly (2007) notes that it may be necessary to hold several short interviews rather than a single long interview, in order for the participants to be able to concentrate, retain their attention (p. 25) and not become tired. She also emphasizes the importance of waiting longer for an answer to be given, and to be alert to different ways in which children can communicate other than through speech (p. 28), for example, facial expression, writing, signing, gestures and non-verbal communication, symbols (see also Mitchell and Sloper, 2008, p. 11), drawing and game playing. There has to be a shift, Kelly avers, away from a deficit model in which children cannot speak, and towards a positive model of how they can communicate through other means.

Morgan (1996) suggests that interviewing marginalized groups can be addressed usefully through group interviewing and with focus groups, and it is to focus groups that we turn now.

25.9 Focus groups

Focus groups are a form of group interview in which reliance is placed on the interaction within the group, which discusses a topic supplied by the researcher (Morgan, 1988, p. 9), yielding a collective rather than an individual view. Here the participants interact with each other rather than with the interviewer, such that the views of the participants can emerge – the participants' rather than the researcher's agenda can predominate. It is from the *interaction* of the group that the data emerge, hence the dynamics of the groups are important (Denscombe, 2014, p. 189). Focus groups are contrived settings, bringing together a specifically chosen sector of the population, often previously unknown to each other (Hydén and Bülow, 2003), to discuss a particular given theme or topic, where the interaction with the group leads to data and outcomes (Smithson, 2000; Hydén and Bülow, 2003). Typically a moderator/facilitator is present to lead the discussion, steer the group as necessary and keep them to the focus of the discussion. The 'contrived' nature of focus groups is both their strength and their weakness: they are unnatural settings yet they are structured and very focused on a particular issue and, therefore, will yield insights that might not otherwise have been gained from a straightforward interview; they are economical on time, often producing a large amount of data in a short period, but they tend to produce less data than interviews with the same number of individuals on a one-to-one basis. Focus groups have the attraction of synergy, with several people stimulating discussion and working together on the issue in hand.

Focus groups (Morgan, 1988; Krueger, 1988; Bailey, 1994, pp. 192–3; Robson, 2002, pp. 284–5; Gibbs, 2012) are useful for:

- orientation to a particular field of focus;
- developing themes, topic and schedules flexibly for subsequent interviews and/or questionnaires;
- generating hypotheses that derive from the insights and data from the group;
- generating and evaluating data from different subgroups of a population;
- gathering qualitative data;
- generating data quickly and at low cost;
- gathering data on attitudes, values, perceptions, viewpoints and opinions;
- empowering participants to speak out, and in their own words;
- encouraging groups, rather than individuals, to voice opinions;
- encouraging non-literate participants;

- providing greater coverage of issues than would be possible in a survey;
- gathering feedback from previous studies.

Focus groups might be useful to triangulate with other forms of interviewing, questionnaire, observation etc. There are several issues to be addressed in running focus groups (Morgan, 1988, pp. 41–8; Gibbs, 2012), for example:

- deciding the number of focus groups for a single topic (one group may be insufficient, as the researcher will be unable to know whether the outcome is unique to the behaviour of the group);
- deciding the size of the group (too small and intra-group dynamics exert a disproportionate effect, too large and the group becomes unwieldy and hard to manage; it fragments). Morgan (1998, p. 43) suggests between four and twelve people per group, whilst Fowler (2009, p. 117) suggests between six and eight people;
- how to allow for people not 'turning up' on the day. Morgan (1998, p. 44) suggests the need to over-recruit by as much as 20 per cent;
- taking extreme care with the sampling, so that *every* participant is the bearer of the particular characteristic required or that the group has homogeneity of background in the required area, otherwise the discussion will lose focus or become unrepresentative; sampling is a major key to the success of focus groups;
- ensuring that participants have something to say and feel comfortable enough to say it;
- chairing the meeting so that a balance is struck between being too directive and veering off the point, i.e. keeping the meeting open-ended but to the point;
- having an effective and well-briefed facilitator to set the ground rules, clarify, probe, question, keep to the point, reflect back, summarize and manage group dynamics etc.;
- how to address confidentiality and informed consent (and indeed other ethical issues), disagreements, conflicts, strong feelings, silence, non-verbal communication and complex responses.

Newby (2010, pp. 350–1) and Gibbs (2012) indicate that focus groups should be clear on the agenda and the focus, take place in a setting that is conducive to discussion, have a skilled moderator (facilitator) who can prompt people to speak, promote thinking and reflection, and should have a record kept.

Unlike group interviewing with children, discussed earlier, focus groups operate more successfully if they are composed of relative strangers rather than friends, unless friendship is an important criterion for the focus (e.g. the group will discuss something that is usually only discussed among friends).

Focus groups are not without their drawbacks. For example, they tend not to yield numerical, quantifiable or generalizable data; the data may be difficult to analyse succinctly; they may yield less information than a survey; the group dynamics may lead to non-participation by some members and dominance by others (e.g. status differentials may operate); the number of topics to be covered may be limited; intra-group disagreement and even conflicts may arise; inarticulate members may be denied a voice; the data may lack overall reliability. Further, Smithson (2000) suggests that there is a problem of only one voice being heard, particularly if there is a dominant member of the group, and for the group dynamics to suppress dissenting voices or different views on controversial topics, even though the group moderator may try to prevent this. Focus groups require skilful facilitation and management by the researcher. Many of the points raised earlier about group interviewing apply to focus groups, and we refer readers to this.

25.10 Non-directive, focused, problem-centred and in-depth interviews

The non-directive interview

Originating from the psychiatric and therapeutic fields with which it is most readily associated, the non-directive interview is characterized as a situation in which the respondent is responsible for initiating and directing the course of the encounter and for the attitudes she expresses in it (in contrast to the structured or research interview, where the dominating role assumed by the interviewer results in differentials of power and commitment (Kitwood, 1977)). It has been shown to be a particularly valuable technique because it reaches the deeper attitudes and perceptions of the person being interviewed in such a way as to leave them free from interviewer bias. We examine here, if briefly, the characteristics of the therapeutic interview and then consider its usefulness as a research tool in the social and educational sciences.

The non-directive interview as it is currently understood grew out of the pioneering work of Freud and subsequent modifications to his approach by later analysts. His basic discovery was that if one can arrange a special set of conditions and have a patient talk about his or her difficulties in a certain way, behaviour changes of many kinds can be accomplished. The

technique developed was used to elicit highly personal data from patients in such a way as to increase their self-awareness and improve their skills in self-analysis. By these means they became better able to help themselves. As Madge (1965) observed, these techniques have greatly influenced interviewing techniques, especially those of a more penetrating and less quantitative kind.

The therapeutic interview has its most persuasive advocate in Carl Rogers, who testified to its efficacy. Basing his analysis on his own clinical studies, he identified a sequence of characteristic stages in the therapeutic process, beginning with the client's decision to seek help. He/she is met by a counsellor who is friendly and receptive, but not didactic. The next stage is signalled when the client begins to give vent to hostile, critical and destructive feelings, which the counsellor accepts, recognizes and clarifies. Subsequently, and invariably, these antagonistic impulses are used up and give way to the first expressions of positive feeling. The counsellor likewise accepts these until suddenly and spontaneously 'insight and self-understanding come bubbling through' (Rogers, 1942, p. 40). With insight comes the realization of possible courses of action and also the power to make decisions. It is in translating these into practical terms that the client frees himself/herself from dependence on the counsellor.

Rogers (1945) subsequently identified a number of qualities in the interviewer which he deemed essential: that she bases her work on attitudes of acceptance and permissiveness; that she respects the client's responsibility for his own situation; that she permits the client to explain his problem in his own way; and that she does nothing that would in any way arouse the client's defences.

What of the usefulness of the non-directive interview as a research technique in educational contexts? There are several features of the therapeutic interview which are peculiar to it and may well be inappropriate in other settings, for example: the interview is initiated by the respondent; his/her motivation is to obtain relief from a particular symptom; the interviewer is primarily a source of help, not a procurer of information; the interview is part of a therapeutic experience; the purpose of the interview is to change the behaviour and inner life of the person and its success is defined in these terms; and there is no restriction on topics discussed.

The focused interview

What appear as advantages in a therapeutic context may be decided limitations when the technique is used for research purposes, even though the interviewer may be sympathetic to the spirit of the non-directive interview. For example, the research interview is less concerned with therapy and 'curing' and having the interviewee control the interview to a great extent, and more concerned with obtaining data. One attempt to meet the need for the research application of the non-directive interview is the *focused interview* (Merton and Kendall, 1946; Flick, 2009). While seeking to follow closely the principle of non-direction, the method introduces rather more interviewer control in the kinds of questions used and seeks also to limit the discussion to certain parts of the respondent's experience.

The focused interview differs from other types of research interview in certain respects (Merton and Kendall, 1946):

- The persons interviewed are known to have been involved in a particular situation: they may, for example, have watched a film; or read a book or article; or have been a participant in a social situation.
- By means of the techniques of content analysis, elements in the situation which the researcher deems significant have previously been analysed by her. She has thus arrived at a set of hypotheses relating to the meaning and effects of the specified elements.
- Using her analysis as a basis, the investigator constructs an interview guide. This identifies the major areas of enquiry and the hypotheses which determine the relevant data to be obtained in the interview.
- The actual interview is focused on the subjective experiences of the people who have been exposed to the situation. Their responses enable the researcher both to test the validity of her hypotheses, and to ascertain unanticipated responses to the situation, thus giving rise to further hypotheses.

The distinctive feature of the focused interview is the prior analysis by the researcher of the situation in which subjects have been involved. The advantages of this procedure have been cogently explained by Merton and Kendall (1946):

> Foreknowledge of the situation obviously reduces the task confronting the investigator, since the interview need not be devoted to discovering the objective nature of the situation. Equipped in advance with a content analysis, the interviewer can readily distinguish the objective facts of the case from the subjective definitions of the situation. He thus becomes alert to the entire field of 'selective

> response'. When the interviewer, through his familiarity with the objective situation, is able to recognize symbolic or functional silences, 'distortions', avoidances, or blockings, he is the more prepared to explore their implications.
>
> (Merton and Kendall, 1946, p. 541)

In the quest for what Merton and Kendall term 'significant data', the interviewer must develop the ability to evaluate continuously the interview while it is in progress. To this end, they established a set of criteria by which productive and unproductive interview material can be distinguished (see also Flick, 2009). Briefly, these are:

- *non-direction*: interviewer guidance should be minimal;
- *specificity*: respondents' definitions of the situation should find full and specific expression;
- *range and scope*: the interview should maximize the range of evocative stimuli and responses reported by the subject;
- *depth and personal context*: the interview should bring out the affective and value-laden implications of the subjects' responses, to determine whether the experience has central or peripheral significance. It should elicit the relevant personal context, the idiosyncratic associations, beliefs and ideas.

The problem-centred interview

Witzel (2000) advocates the use of the *problem-centred interview* for gathering objective evidence on human behaviour and subjective views on social phenomena. He indicates three principles underpinning the problem-centred interview:

- a 'problem-centred orientation' towards socially relevant problems (p. 2);
- methodological flexibility (e.g. group interviews, individual interviews, the biographical interview, structured and less structured interviews) in the 'object-orientation', in order to address different kinds of problem (p. 3);
- a 'process orientation', i.e. attempting to reconstruct the actions and orientations of the participant (p. 3).

The problem-centred interview, Witzel avers, can use: (a) a structured, *short questionnaire* at interview in order to gather factual data about the participants (e.g. age, sex, occupation, education); (b) an *interview schedule* (guidelines) in order to structure the interview, with lead questions; (c) *recording equipment* to ensure accuracy of the account and to avoid the interviewer having to take notes (i.e. able to concentrate on the discussion); and (d) a *postscript* (pp. 3–4), written directly after the interview, to contain reflections, key points, observations and interpretations. Flick (2009), like Witzel, advocates the preparation of an interview guide in the problem-centred interview, consideration of how to manage 'conversational entry' (p. 163) prompting (general and specific: deep) and being prepared to raise ad hoc questions and to challenge the interviewee on contradictions and inconsistencies in statements made.

The in-depth interview

The in-depth interview, as its name suggests, is conducted to explore issues, personal biographies, what is meaningful to, or valued by, participants, what they know about a topic, what they have experienced, how they feel about particular issues, how they look at particular issues, their attitudes, opinions and emotions (cf. Newby, 2010, pp. 243–4; Mears, 2012). They tend to be semi-structured, to enable the course of the respondents' responses to dictate the direction of the interview, though the researcher also has an interview schedule to keep an interview on track, and may operate probes to inquire further into issues. They may feature in case studies, action research and, as the work of Ball (1990, 1994a, 1994b), Bowe *et al*. (1992) and Flick (2009) testifies, they may feature in interviewing powerful people and policy makers. They may also be useful in gathering data from marginalized or stigmatized groups in society (Newby, 2010, pp. 243–4). Given the intensive and extensive nature of these interviews, gaining access and permission may be difficult, not least as they may take time to conduct.

25.11 Telephone interviewing

We advise the reader to take this section in conjunction with Chapter 17 on telephone interviews in surveys. The use of telephone interviewing has long been recognized as an important method of data collection and is common practice in survey research. Arksey and Knight (1999, p. 79) note that telephone interviews do not feel like interviews, as both parties are deprived of several channels of communication and the establishment of a positive relationship (e.g. non-verbal). Similarly, Lechuga (2012) notes that telephone interviews are often used in tandem with face-to-face interviews, though in survey research they can be the principal means of data collection.

Dicker and Gilbert (1988), Nias (1991), Oppenheim (1992), Borg and Gall (1996), Shaughnessy *et al*. (2003), Shuy (2003), Lechuga (2012) and Kee and

Browning (2013) suggest several attractions to telephone interviewing:

- it is sometimes cheaper and quicker than face-to-face interviewing;
- it enables researchers to select respondents from a much more dispersed population than if they have to travel to meet the interviewees;
- travel costs are omitted;
- interviews can be conducted at a time to suit the interviewee, and can reduce interruptions;
- it is particularly useful for brief surveys;
- it can protect the privacy, anonymity and confidentiality of respondents, and this can improve the quality of responses, according them safety;
- it is useful for gaining rapid responses to a structured questionnaire;
- monitoring and quality control are undertaken more easily since interviews are undertaken and administered centrally, indeed there are greater guarantees that the researcher actually carries out the interview as required;
- interviewer effects are reduced;
- there is greater uniformity in the conduct of the interview and the standardization of questions;
- there is greater interviewer control of the interview;
- the results tend to be quantitative (e.g. with standardized interview questions), giving a degree of consistency/control to the interview;
- telephone interviews are quicker to administer than face-to-face interviews because respondents will only usually be prepared to speak on the telephone for, at most, ten to fifteen minutes;
- call-back costs are so slight as to enable frequent call-backs, enhancing reliability and contact;
- many groups, particularly of busy people, can be reached at times more convenient to them than if a visit were to be made;
- it can reach marginalized or minority groups;
- it can neutralize power differentials between interviewer and interviewee;
- it can increase disclosure by interviewees in comparison to face-to-face interviews, and can be used to collect sensitive data, as possible feelings of threat of face-to-face questions about awkward, embarrassing or difficult matters are absent;
- it is safer to undertake than, for example, having to visit dangerous neighbourhoods;
- it does not rely on the literacy of the respondent (as, for example, in questionnaires);
- it may put a little pressure on the respondent to respond, and it is usually the interviewer rather than the interviewee who terminates the call;
- response rate is higher than, for example, questionnaires.

Smartphones are useful in interviewing (Raento *et al*., 2009), as so many people carry them around and hence can be contacted easily, regardless of their location (though it relies on the interviewer knowing the smartphone number). Further, with increasing additional functionality, the smartphone offers a potentially powerful tool for researchers. Smartphones, Raento *et al*. aver (2009, p. 428), have the attraction of flexibility, with many uses and computer-like functions on a single smartphone, for example, video and image recording and cost-efficiency; they enable easy access and relatively unobtrusive data collection as data can be collected in real time or stored; and they enable high granularity of data to be gathered (p. 442). We refer the reader to the discussion of online interviews later in this chapter.

Telephone interviewing is not as cut-and-dried as the claims made for it, as there are several potential problems with it, for example (e.g. Lechuga, 2012):

- it is very easy for respondents simply to hang up on the caller;
- motivation to participate may be lower than for a personal interview;
- there is a chance of skewed sampling, as not all of the population have a telephone (perhaps the very people that the researcher wishes to target) or can hear (e.g. the old and second-language-speakers in addition to those with hearing difficulties);
- there is a lower response rate at weekends;
- the lack of visual, non-verbal and contextual cues and data may cause problems for both parties;
- the standardized format of telephone interviews may prevent thoughtful or deep answers from being provided;
- some people may have a deep dislike of telephones that sometimes extends to a phobia, and this inhibits their responses or willingness to participate;
- respondents may not disclose information because of uncertainty about actual (even though promised) confidentiality;
- respondents may come to snap judgements without the adequate or deeper reflection necessary for a full answer to serious issues;
- respondents may not wish to spend a long time on the telephone, so telephone interviews tend to be briefer than other forms of interview;
- concentration spans are shorter than in a face-to-face interview;
- the interviewer has to remain bright and focused, listen very carefully and respond, which is tiring;

- questions tend to be closed, fixed and simple and there is a limit on the complexity of the questions that can be put;
- the quality of the data may be inferior than face-to-face interview data because of shortage of time, lack of visual and non-verbal cues and limited emotional feedback;
- the limits placed on emotional feedback from the interviewer may prevent the interviewee from disclosing sensitive information;
- response categories must be very simple or else respondents will forget what they are;
- many respondents may be 'ex-directory' or numbers will not be available (e.g. in telephone directories), particularly for cellphones;
- respondents may withhold important information or tell lies, as the non-verbal behaviour that frequently accompanies this is not witnessed by the interviewer;
- it is often more difficult for complete strangers to communicate by telephone than face-to-face, particularly as non-verbal cues are absent;
- respondents are naturally suspicious (e.g. of the caller trying to sell a product);
- one telephone might be shared by several people;
- some telephone service providers screen out callers;
- some respondents feel that telephone interviews afford less opportunity for interviewees to question or rebut the points made by the interviewer;
- there may be distractions for the respondent (e.g. a television may be switched on in the background, children may be crying, others may be present);
- responses are difficult to write down or record during the interview.

Harvey (1988), Oppenheim (1992), Miller (1995) and Lechuga (2012) consider that: (a) telephone interviews need careful arrangements for timing and duration (as they are typically shorter and quicker than face-to-face interviews) – a preliminary call may be necessary to fix a time for a longer call to be made; (b) the interviewer will need to have ready careful prompts and probes, including more than usual closed questions and less complex questions, in case the respondent 'dries up' on the telephone; (c) both interviewer and interviewee need to be prepared in advance of the interview if its potential is to be realized; and (d) sampling requires careful consideration, using, for example, random sampling or some form of stratified sample. In general, however, many of the issues from 'standard' forms of interviewing apply equally well to telephone interviewing. Further, Houtkoop-Steenstra and van den Bergh (2000) report that an agenda-based introduction (in which interviewers formulate their own introductions based on a small number of keywords) is more effective in securing higher response rates than standardized, scripted introductions.

Kee and Browning (2013, pp. 12–16) note that interviewers can use different ways to persuade interviewees to take part in a telephone interview, for example: persuasion by association (at the end of one interview the interviewer asks the interviewee to recommend other people for interview); persuasion by specificity (being specific on what the interview will cover); persuasion by trust: the interviewer assures the interviewee that the data will not be abused or misused or cause problems for the interviewee (i.e. non-maleficence); persuasion by kindness/goodwill (interviewees agree to be kind and helpful); persuasion by opportunity (e.g. to take part in an important piece of research); persuasion by personalization (a personal invitation); persuasion by flexibility (the interviewer can call at the interviewee's convenience); and persuasion by sequence (an email is followed by the telephone call).

Face-to-face interviews may be more suitable than telephone interviews (Weisberg *et al*., 1996, p. 122; Shuy, 2003, pp. 179–82; Kee and Browning, 2013) where: (a) the interviewer wishes to address complex issues or sensitive questions (though these authors also note that telephone interviewing may be particularly useful for exploring sensitive issues because of the lack of a face-to-face presence); (b) a natural context might yield greater accuracy; (c) deeper and self-generated answers are sought, i.e. where the question does not frame the answer too strongly; (d) issues requiring probing, deep reflection and, thereby, a longer time is sought; (e) greater equality of power between interviewer and respondent is sought; (f) older, second-language-speakers and hearing-impaired respondents are being interviewed; (g) marginalized respondents are being sought.

In conducting the telephone interview, the interviewer should strive to put the participant at ease, introduce himself/herself and to clarify any uncertainties (Kee and Browning, 2013), obtain consent and build trust. It may also be necessary for the interviewer to clarify ('customize') pre-scripted questions (p. 20). Further, in a telephone interview, as in face-to-face interviews, the interviewer's personality and skill are important factors in success, as both parties – interviewer and interviewee – are striving to define the attitudes, nature and personality of the other (Lechuga, 2012).

It is not uncommon for telephone interviewing to be outsourced; this can be an advantage or a disadvantage. On the one hand, it takes pressure off the researcher, not only because of the time involved but also because a fifteen-minute telephone interview might be more

exhausting than a fifteen-minute face-to-face interview, there being more social and non-verbal cues in face-to-face interaction. On the other hand, in outsourced telephone interviews care has to be taken on standardization of the conduct of the interview, the content of questions, the entry of responses, the personality and demeanour of the interviewer, and indeed to check that the interviews have been done and responses not simply fabricated.

In conducting telephone interviews it is important to consider several issues:

- Will the people have the information that you require? Who will you need to speak to on the telephone? If the person answering the call is not the most suitable person then you need to talk to somebody else.
- There is a need to pilot the interview schedule and to prepare and train the telephonists, to discover the difficult/sensitive/annoying/personal questions, the questions over which the respondents hesitate and answer very easily, the questions that will need prompts and explanations.
- Keep to the same, simple response categories for several questions, so that the respondents become used to these and keep in the same mindset for responding.
- Keep personal details, if any, until the end of the interview, in order to reduce any sense of threat.
- Keep to no more than, at the most, thirty questions, and take no longer than, at the most, fifteen minutes, and preferably ten minutes.
- Be clear with the respondents at the start of the interview that they have the time to answer and that they have the information sought (i.e. that they are suitable respondents). If they are not the most suitable respondents, then ask if there is someone present who can answer the questions, or try to arrange call-back times when the most suitable person can be reached. Ask to speak to the most suitable person.
- Keep the terminology simple and to the point, avoiding jargon and confusion.
- You should be able to tell the gender of the respondent by his or her voice, i.e. there may be no need to ask a particular question.
- Keep the response categories very simple and use them consistently (e.g. a mark out of ten, 'strongly agree' to 'strongly disagree', a 1–5 scale etc.).
- Rather than asking direct personal questions (unless you are confident of an answer), for example, about age or income, ask about categories such as which age group or income group they fall into (and give the age groups or income brackets).

25.12 Online interviewing

Online interviewing, interviewing in the virtual world, takes several forms, for example:

- text-based only (chat rooms, social networking sites, discussion forums, blogs, email, SMS);
- a combination of text and visuals (e.g. social networking sites, discussion boards, instant messaging);
- audio only (e.g. WeChat, WhatsApp, conferencing);
- audio and visual interviews (e.g. Skype, net meetings and conferences);

Some of these are synchronous – real-time – messaging (e.g. chat rooms and services, SMS, instant messaging, Skype, online meetings); others are asynchronous (e.g. email, discussion boards, blogs, WeChat, WhatsApp). Some online interviews take place through private contact (e.g. emails), others through public sites (e.g. blogs), and researchers need to consider the issue of privacy.

What they have in common is that they all involve versions of online interviewing. Different kinds have different advantages and disadvantages; we explore these below and we also advise readers to review Chapter 18. However, all of these presume access to the Internet, smartphone or relevant equipment/hardware, and the ability to download and use relevant software, and these latter might be problematic, even for more experienced and competent IT users. Researchers are advised to avoid uncommonly used, difficult-to-use software and software which requires complicated registration, and to opt for commonly used software.

A great attraction of these methods is that time and location present few or no challenges, as there is great flexibility in contact times, and both parties do not necessarily have to be present at the same time for some of these methods. Contact can be made with complete strangers, anonymity can be preserved in some of these methods, and, for non-visual methods, the absence of a face might encourage more sensitive areas to be addressed (as with the comments earlier about telephone interviewing). Further, the potential differentials of power between interviewer and interviewee can be reduced (James, 2015, 2016).

In synchronous, text-based interviews, the speed of the interview can depend on speed of connection, typing and reading, whilst in asynchronous methods the interviewers have the flexibility and convenience of different contact times and response times, the opportunity to think about, and reflect on, questions and previous answers, and to consider carefully their answers.

Some text-based methods can be both synchronous and asynchronous, depending on the wishes and availability of the users (e.g. email, chat rooms, SMS).

On the other hand, the interviewer has no control over the circumstances of the interviewee, who may be distracted, indeed who may not be the person intended – an identity problem – and who may not have the level of motivation or interest to participate (and it is often easier to decline an online interview than a face-to-face interview). Some interviewees may not wish to read/write/type in text-based interviews, or be less competent in doing so, and text-based interviews deprive both parties of the benefits of visual clues and non-verbal communication (as with telephone interviewing). Additionally, research using online interviews, particularly if they are text-based only, must avoid 'flaming' (overreacting with aggressive, insulting, attacking, derogatory or hostile remarks).

The considerations for interviews already set out in this chapter and Chapters 12, 17 and 18 apply to online interviews, for example:

- sampling and representativeness;
- whether to have individual or group interviews;
- the need to have interview questions prepared;
- the need to plan the content, structure, sequence and type of questions;
- the need to plan and consider probes and follow-ups;
- the need to build relationships, rapport and trust;
- how to overcome lack of visual/non-verbal cues;
- the need to observe research and interview ethics;
- the need to avoid risks of bias and socially desirable answers;
- the need to choose language, genre and register very carefully.

James and Busher (2007) and James (2007, 2015, 2016) writing about emails, and Hinchcliffe and Gavin (2008), Kazmer and Xie (2008) and Pearce *et al*. (2014) writing about instant messaging, argue for the power of email/messaging interviewing as a qualitative method in educational research, as it enables the researcher to contact hard-to-reach groups or individuals, for example, by virtue of practical constraints such as time and availability of both parties to meet face-to-face, location and travelling, geographical dispersion, disability and language or communication (James, 2007, 2015, 2016; see also Bampton and Cowton, 2002).

Whilst email typically has a time delay (which can be very short indeed: a matter of a few seconds), instant messaging is, as its name suggests, instant, i.e. synchronous. Instant messaging can also be aural. Pearce *et al*. (2014) suggest that instant messaging (which now includes visual functions) trumps email for research purposes even though it requires prior arrangements to be made for the interview to be synchronous. However, James and Busher (2007, p. 405) argue that email interviews, by virtue of their ability to be non-synchronous, can generate fuller, richer, more reflective, thoughtful and longer answers than telephone interviewing, and James (2016) notes that emails afford the researcher the opportunity to probe the interviewee for further or richer information.

Email interviews also reduce transcription time as the email is already transcribed, and the interviewee, therefore, has the opportunity to check what data are being given, thereby overcoming the possibility of respondents in a face-to-face interview saying something they later wish to withdraw (Bampton and Cowton, 2002, p. 4). Whether or not the absence of face-to-face contact, the visibility of the participants and the absence of non-verbal cues increases or reduces reliability is a moot point (James, 2007, p. 969). Similarly, the researcher will need to make continual efforts to ensure that the respondent:

- keeps to the point;
- fully understands the nature, focus and purpose of the interview;
- knows the number of questions that will be asked (particularly if there are several email exchanges);
- knows that they should not delete previous emails that are part of the interview;
- knows the time frame in which to reply to an email (cf. James and Busher, 2006, pp. 407–9).

Email interviews can be conducted synchronously, in real time, or asynchronously (James, 2007, 2015, 2016). The latter can afford the interviewee some time to consider his/her responses (Bampton and Cowton, 2002, p. 3), and the 'silences' (James, 2015, 2016) between question, answer and the subsequent question/follow-up can be useful pauses to think, even when such silences span week or months. Email interviewing can 'democratise narrative exchanges' – as equals – between the interviewer and the interviewee (James, 2007, p. 970) (though James and Busher (2007, p. 416) contest this latter point). However, Bampton and Cowton (2002, p. 5) caution against bombarding the interviewee with too many questions in a single interview; rather, they suggest, the questions could be spaced over more than one email. Further, they indicate the need to signal to the respondent when the email interview is nearing its close.

Online interviewing is susceptible to technological problems (e.g. unstable connectivity, slow connections (particularly in video-conferencing), mailbox being full), and these must be explored before the online interview is conducted.

Skype, as an instant, face-to-face audio-visual method, has all the benefits and drawbacks of the face-to-face interview, and it can also include instantaneous text-based messaging. Both the researcher and interviewee have to agree the time for the interview, and each party might not know if there are any other parties present in the location.

Researchers will need to consider whether it is preferable to use text-based-only methods for online interviewing, for example, with the advantages of 'opaqueness', anonymity and confidentiality that might be highly suited to sensitive topics and provide the opportunity for reflection and review (Pearce *et al*., 2014) or whether to include visual methods which have the attractions of seeing body language and facial response. In both cases there are risks of bias. This is akin to the argument for and against face-to-face versus telephone interviewing, played out this time on a virtual plane.

We sum up the different forms of administering interviews in Figure 25.1.

25.13 Ethical issues in interviewing

Interviews have an ethical dimension; they concern interpersonal interaction and produce information about the human condition. Though one can identify several main areas of ethical issues here – informed consent, beneficence, do no harm, confidentiality and the consequences of the interviews – these need to be 'unpacked' a little, as each is problematic (Kvale, 1996, pp. 111–20). For instance, who should give the informed consent (e.g. participants, their superiors), and for whom and what? How much information should be given, and to whom? What is legitimate private and public knowledge? How might the research help or harm the interviewees? Does the interviewer have a duty to point out the possible harmful consequences of the research data or will this illegitimately steer the interview? How will the interviewers and interviewees both benefit from/gain something from the interview (cf. Mills, 2001)?

It is difficult to lay down hard and fast ethical rules, as ethical matters are situated, contestable and context-based. Nevertheless it is possible to raise some ethical questions to which answers need to be given before the interviews commence:

- Has the informed consent of the interviewees been gained?
- Has this been obtained in writing or orally?
- How much information should be given in advance of the study?
- How can adequate information be provided if the study is exploratory?
- Have the possible consequences of the research been made clear to the participants?
- Has care been taken to prevent any harmful effects of the research to the participants (and to others)?
- To what extent do any potential benefits outweigh the potential harm done by the research, and how justifiable is this for conducting the research?
- How will the research benefit the participants?
- Who will benefit from the research?
- How much reciprocity is there between what participants give to and receive from the research?
- Have confidentiality, anonymity, non-identifiability and non-traceability been guaranteed? Should participants' identities be disguised?
- How do Data Protection Acts and laws operate in interview situations?
- Who will have access to the data?
- What has been done to ensure that the interview is conducted in an appropriate, non-stressful, non-threatening manner?
- How will the data and transcriptions be verified, and by whom?

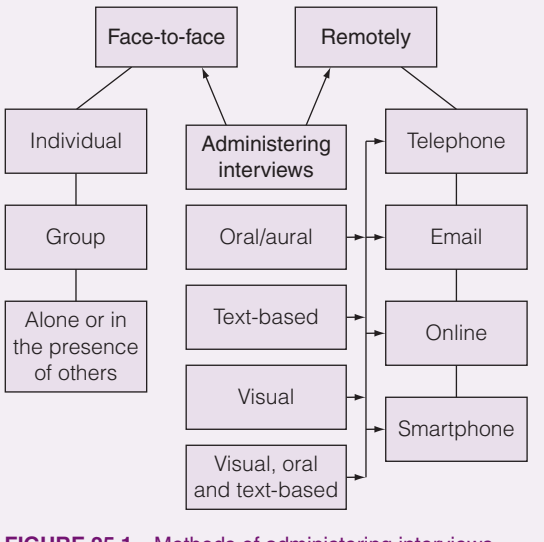


FIGURE 25.1 Methods of administering interviews

- Who will see the results of the research? Will some parts be withheld? Who owns the data? At what stage does ownership of the data pass from interviewees to interviewers? Are there rights of veto for what appears? To whom should sensitive data be made available (e.g. should interview data on child abuse or drug-taking be made available with or without consent of parents and the police)?
- How far should the researcher's own agenda and views predominate? What if the researcher makes a different interpretation from the interviewee? Should the interviewees be told, even if they have not asked for these interpretations?

These issues, by no means an exhaustive list, are not exclusive to the research interview, though they are applicable here. For further reading on ethical issues, we refer readers to Chapters 7 and 8. The personal safety of interviewers must also be addressed: it may be important, for example, for the interviewer to be accompanied, to leave details of where he or she is going, to take a friend, to show identification, to take a mobile phone, to reconnoitre the neighbourhood, to learn how to behave with fierce dogs or people, to use the most suitable transport. It is perhaps a sad indictment on society that these considerations have to be addressed, but they do.

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Observation

CHAPTER 26

Observation takes many forms. This chapter addresses several different kinds of observation and how to plan, conduct and report them, and we address:

- structured observation
- the need to practise structured observation
- analysing data from structured observations
- critical incidents
- naturalistic and participant observation
- data analysis for unstructured observations
- natural and artificial settings for observation
- video observations
- timing and causality with observational data
- ethical considerations
- reliability and validity in observations

This chapter can also be read with Chapters 15 and 31.

26.1 Introduction

Observation is more than just looking. It is looking (often systematically) and noting systematically people, events, behaviours, settings, artefacts, routines, and so on (Simpson and Tuson, 2003, p. 2; Marshall and Rossman, 2016). It can be systematic and structured or take some less structured form such as participant observation (Denscombe, 2014, p. 205).

The distinctive feature of observation as a research process is that it offers an investigator the opportunity to gather first-hand, 'live' data *in situ* from naturally occurring social situations rather than, for example, reported data (Wellington, 2015, p. 247) and second-hand accounts (Creswell, 2012, p. 213). The use of observation as a principal mode of research has the potential to yield more valid or authentic data than would otherwise be the case with mediated or inferential methods. This is observation's unique strength. Observation is strong on face validity; it can provide rich contextual information, enable first-hand data to be collected, reveal mundane routines and activities, and can offer an opportunity for documenting those aspects of lifeworlds that are verbal, non-verbal and physical (Clark *et al*., 2009).

There are other attractions to observation: as Robson says (2002, p. 310), what people do may differ from what they say they do, and observation provides a reality check. Observation also enables a researcher to look afresh at everyday behaviour that otherwise might be taken for granted, expected or go unnoticed (Cooper and Schindler, 2001, p. 374), and the approach, with its carefully prepared recording schedules, avoids problems such as selective or faulty memory caused when there is a time gap between the act of observation and the recording of the event. On a procedural point, some participants may prefer the presence of an observer to an intrusive, time-consuming interview or questionnaire.

Observation can be of *facts*, for example: the number of books in a classroom; the number of students who visit the school library in a given period. It can also focus on *events* as they happen in a classroom, for example: the amount of teacher and student talk; the amount of off-task conversation. It can focus on *behaviours* or qualities, for example: the friendliness of the teacher; the extent of cooperative behaviour among students.

There is a continuum from the observation of uncontestable facts to the researcher's interpretation and judgement of situations, which are then recorded as observations. What counts as evidence immediately becomes cloudy in observation, because what we observe depends on when, where and for how long we look, how many observers there are and how we look. Indeed the post-positivists argue that there is no neutral, theory-free observation and that our observations are driven by our theories (conscious or not, implicit or explicit), experiences, prior concepts, preferences, values, beliefs, purposes, assumptions, foci, perceptions and intentions (Denscombe, 2014, p. 206; Pring, 2015, p. 48; Wellington, 2015, p. 248). Observations are theory-laden and experience-laden (Barrett and Mills, 2009). Whilst reflexivity tries to address this, it only exposes rather than solves the problem.

What we observe also depends on what is taken to be evidence of, or a proxy for, an underlying, latent construct. What counts as acceptable evidence requires an operational definition that is valid and reliable.

Observers need to decide 'what is the observation evidence', for example: is the degree of wear and tear on a book in the school library an indication of its popularity, or carelessness by its readers, or of destructive behaviour by students? It is dangerous to infer cause from effect, intention from observation, stimulus from response.

Observational data are sensitive to context and require strong ecological validity (Moyles, 2002) to understand the context of programmes, to be open-ended and inductive, to see things that might otherwise be unconsciously missed, to discover things that participants might not freely talk about in interview situations, to move beyond perception-based data (e.g. opinions in interviews) and to access personal knowledge. Because observed incidents are less predictable there is freshness in this form of data collection that is often denied in other forms, for example, a questionnaire or a test.

Observation can enable the researcher to access interactions in a social context and to yield systematic records of these in many forms and contexts, to complement other kinds of data (Simpson and Tuson, 2003, p. 17). Observations (Morrison, 1993, p. 80) enable the researcher to gather data on:

- the *physical setting* (e.g. the physical environment and its organization);
- the *human setting* (e.g. the organization of people, the characteristics and make-up of the groups or individuals being observed, for instance, gender, class);
- the *interactional setting* (e.g. interactions that are taking place, formal, informal, planned, unplanned, verbal, non-verbal etc.);
- the *programme setting* (e.g. resources and their organization, pedagogic styles, curricula and their organization).

Additionally, observational data may be useful for recording non-verbal behaviour, behaviour in natural or contrived settings, and longitudinal analysis (Bailey, 1994, p. 244). On the other hand, the lack of control in observing in natural settings may render observation less useful, coupled with difficulties in measurement, problems of small samples, difficulties of gaining access and negotiating entry, and difficulties in maintaining anonymity (pp. 245–6). Observation can be a powerful research tool, but it is not without its difficulties, and this chapter identifies and addresses these.

Patton (1990, p. 202) suggests that observational data should enable the researcher to enter and understand the situation that is being described. The kind of observations available to the researcher lies on a continuum from unstructured to structured, responsive to pre-ordinate. A *highly structured* observation will know in advance what it is looking for (i.e. pre-ordinate observation) and will have its observation categories worked out in advance (e.g. Heath *et al*., 2010). A *semi-structured observation* will have an agenda of issues but will gather data to illuminate these issues in a far less predetermined or systematic manner (e.g. responsive to what is observed). An *unstructured observation* will be far less clear on what it is looking for and will therefore have to go into a situation and observe what is taking place before deciding on its significance for the research. In a nutshell, a structured observation will already have its hypotheses decided in advance and will use the observational data to confirm or refute these hypotheses. On the other hand, a semi-structured and, more particularly, an unstructured observation will be hypothesis-generating rather than hypothesis-testing. Semi-structured and unstructured observations will review observational data before suggesting an explanation for the phenomena being observed.

There is a well-known classification of researcher roles in observation, which lie on a continuum:

- the *complete participant*: a member of the group who conceals her/his role as an observer, whose knowledge of the group/situation may be intimate and who may gain 'insider knowledge', but who may be viewed with suspicion or resentment by the other members when his/her true role comes to light and who may lack the necessary objectivity to observe reliably;
- the *participant-as-observer*: a member of the group who reveals her/his role as an observer, whose knowledge of the group/situation may be intimate and who may gain 'insider knowledge', but who may lack the necessary objectivity to observe reliably and with whom confidences and confidential data may not be shared or given respectively;
- the *observer-as-participant*: not a member of the group, but who may participate a little or peripherally in the group's activities, and whose role as researcher is clear and overt, as unobtrusive as possible, without those being observed always knowing who is the researcher, and whose access to information and people may be incomplete or restricted;
- the *complete observer*: who only observes (overt or covert) and is detached from the group, for example, an outside observer, or where the observer is not covert but whose presence is unnoticed by the group, for example, an observer in a crowded location.

The move is from complete participation to complete detachment. The mid-points of this continuum strive to balance involvement with detachment, closeness with distance, familiarity with strangeness. The role of the complete observer is typified in the two-way mirror, the video recording and the photograph, whilst complete participation involves researchers taking on membership roles (overt or covert).

Traditionally, observation has been characterized as non-interventionist (Adler and Adler, 1994, p. 378), where researchers do not seek to manipulate the situation or subjects, they do not pose questions for the subjects, nor do they deliberately create 'new provocations' (p. 378). Quantitative research tends to have a small field of focus, fragmenting the observed into minute chunks that can subsequently be aggregated into a variable. Qualitative research, on the other hand, draws the researcher into the phenomenological complexity of participants' worlds; here situations unfold, and connections, causes and correlations can be observed as they occur over time. The qualitative researcher aims to catch the dynamic nature of events, to see intentionality and maybe to seek trends and patterns over time.

If we know in advance what we wish to observe, if the observation is concerned to chart the *incidence*, *presence* and *frequency* of elements and wishes to compare one situation with another, then it may be efficient in terms of time to go into a situation with a prepared observation schedule. If, on the other hand, we want to go into a situation and let the elements of the situation speak for themselves, perhaps with no concern for how one situation compares with another, then it may be more appropriate to opt for a less structured observation.

The former, structured observation, can take much time to prepare but the data analysis is fairly rapid, the categories having already been established, whilst the latter, less structured approach, is quicker to prepare but the data take much longer to analyse. The former approach operates within the agenda of the researcher and hence might neglect aspects of settings if they do not appear on the observation schedule, i.e. it looks selectively at situations. By contrast, the latter operates within the agenda of the participants, i.e. it is responsive to what it finds and therefore, by definition, is faithful to the situation as it unfolds. Here selectivity derives from the *situation* rather than from the *researcher* in the sense that the key issues which emerge follow from the observation rather than the researcher knowing in advance what those key issues will be. Structured observation is useful for testing hypotheses, whilst unstructured observation provides a rich description of a situation which, in turn, can lead to the subsequent generation and testing of hypotheses.

Flick (1998, p. 137) suggests that observation can be considered along five dimensions:

- structured, systematic and quantitative observation versus unstructured and unsystematic and qualitative observation;
- participant observation versus non-participant observation;
- overt versus covert observation;
- observation in natural settings versus observation in unnatural, artificial settings (e.g. a 'laboratory' or contrived situation);
- self-observation versus observation of others.

Cooper and Schindler (2001, p. 375) suggest that observation can be considered along three dimensions:

- whether the observation is direct or indirect (the former requiring the presence of the observer; the latter requiring recording devices, e.g. video cameras);
- whether the presence of the observer is known or unknown (overt or covert research, whether the researcher is concealed (e.g. through a two-way mirror or hidden camera) or partially concealed, i.e. the researcher is seen but not known to be a researcher, e.g. the researcher takes up a visible role in the school);
- the role taken by the observer (participant to non-participant observation, discussed below).

We address these throughout the chapter, and present these dimensions and others in Figure 26.1.

Observation, in general, is not only time-consuming but prone to bias in terms of what, why, when, where, whom and how the observer is observing. Observations are inevitably selective, and, in part, depend as much on the observer's attention and opportunity to observe as they do on the observational instruments and data-collection techniques used. Hence great caution and reflexivity are requisites for this form of data collection. As with other forms of data collection, observational data must enable the research questions to be answered. Indeed Simpson and Tuson (2003, chapter 2) and Hamilton and Corbett-Whittier (2013, p. 99) suggest that observers need to consider:

- the focus of the observation(s);
- why they are observing (e.g. looking for regularities, similarities, evolution of a situation, irregularities, patterns, key features etc.);

FIGURE 26.1 Continua of observation

- the research questions that the observational data will address;
- the boundaries of the observation (what to include and exclude);
- how to record the observations;
- what observer role to take;
- where to observe (e.g. key social places);
- what to observe (e.g. significant objects, setting; events, people etc.);
- whom to observe (e.g. key people, everyday participants, marginalized people);
- how many people, events, settings to observe (i.e. sampling);
- how many observations and series of observations to make;
- how systematic, structured, descriptive to be;
- what is the 'unit' of observation (e.g. a teacher, a student, a pair, a small group, a class);
- what resources are necessary (e.g. human observers, video cameras);
- problems that might be encountered;
- additional information that may be needed to complement the observational record;
- the processing and analysis of observational data;
- how to link observations with other kinds of data.

There is also the need to consider who the observer will be, as observation (particularly participant observation) can be affected by the gender, ethnicity, class, appearance, age, language, personality, temperament, attitude, interpersonal behaviour, familiarity with the situation, involvement and concern etc. of the observer (cf. Kawulich, 2005, p. 7).

On a practical level the researcher has to decide fundamental matters such as whether to stand or sit, whether to move around a setting (e.g. in order to track a student), and where to stand or sit. If a researcher is located too close he/she might be intrusive or inhibiting, or the researcher might lose the observation if a student moves away; if the researcher is too far away he/she might miss the detail of what is happening (cf. Simpson and Tuson, 2003, pp. 54–5).

26.2 Structured observation

A structured, systematic observation enables the researcher to generate numerical data from the observations. Numerical data, in turn, facilitate the making of comparisons between settings and situations, and enable frequencies, patterns and trends to be noted or calculated. The observer adopts a passive, non-intrusive role, simply noting down the incidence of the factors being studied. Observations are entered onto an observational schedule. An example of this is shown in Table 26.1. This is an example of a schedule used to monitor student and teacher conversations over a ten-minute period. The upper seven categories indicate who is speaking to whom, whilst the lower four categories indicate the nature of the talk. Looking at the example of the observation schedule, several points can be noted:

- the categories for the observation are discrete, i.e. there is no overlap between them; for this to be the case a pilot has to be developed and tested in order to iron out any problems of overlap of categories;
- each column represents a 30-second time interval, i.e. the movement from left to right represents the chronology of the sequence, and the researcher has to enter data in the appropriate cell of the matrix every thirty seconds (see section below on 'instantaneous sampling');
- because there are so many categories which must be scanned at speed (every thirty seconds), the

TABLE 26.1 A STRUCTURED OBSERVATION SCHEDULE

Student to Student	/	/	/	/															
Student to Students					/	/													
Student to Teacher											/	/	/	/					
Students to Teacher							/	/	/	/	/								
Teacher to Student															/	/			
Teacher to Students																	/	/	/
Student to Self																			
Task in hand					✓	✓					✓	✓	✓	✓	✓	✓	✓	✓	✓
Previous task						✓	✓	✓	✓	✓	✓								
Future task																			
Non-task	✓	✓	✓	✓															

Notes
/ = participants in the conversation
✓ = nature of the conversation.

researcher will need to practise completing the schedule until he or she becomes proficient and consistent in entering data so that the observed behaviours, settings etc. are entered into the same categories consistently, achieving reliability. This can be done either through practising with video material or through practising in a live situation with participants who will not subsequently be included in the research. If there is to be more than one researcher then it may be necessary to provide training sessions so that the team of researchers proficiently, efficiently and consistently enter the same sort of data in the same categories, i.e. that there is inter-rater reliability;

- the researcher will need to decide what entry is to be made in the appropriate category, for example: a tick (✓), a forward slash (/), a backward slash (\), a figure (1, 2, 3 etc.), a letter (a, b, c etc.), a tally mark (|). Whatever code or set of codes is used, it must be understood by all the researchers (if there is a team) and must be simple and quick to enter (i.e. symbols rather than words), and the researcher will need to become proficient in fast and accurate data entry of the appropriate codes.

The need to pilot a structured observation schedule cannot be overemphasized. Categories must be mutually exclusive, comprehensive, complete, relevant, observable, unambiguous, self-evident and easy to record (cf. Denscombe, 2014, p. 208; Webster, 2015, p. 994). The researcher, then, will need to decide:

- the foci of the observation (e.g. people, events etc.);
- the frequency of the observations (e.g. every thirty seconds, every minute, every two minutes);
- the length of the observation period (e.g. one hour, twenty minutes);
- what counts as evidence (e.g. how a behaviour is defined and operationalized);
- the nature of the entry (the coding system).

The criterion of 'fitness for purpose' is used for making decisions on these five matters. Structured observation will take much time in preparation but the analysis of the data should be rapid as the categories for analysis will have been built into the schedule itself. If close, detailed scrutiny is required then the time intervals will be very short, and if less detail is required then the intervals may be longer.

Dyer (1995, pp. 181–4) suggests that structured observation must address several key matters:

- the choice of the environment (such that there will be opportunities for the behaviour to be observed to actually occur – the availability and frequency of the behaviour of interest to the observer: a key feature if unusual or special behaviour is sought);
- the need for clear and unambiguous measures and proxy measures (particularly if a latent characteristic or construct is being operationalized);
- a manageable number of variables (a sufficient number for validity to be demonstrated, yet not so many as to render data entry unreliable);
- overt or covert observation;
- continuous, time series or random observation;
- the different categories of behaviour to be observed;

- the number of people to be observed;
- the number of variables on which data must be gathered;
- the kind of observation schedule to be used.

Dyer (1995, p. 186) provides a checklist for planning a structured observation, shown in Box 26.1.

We introduce here five principal ways of entering data onto a structured observation schedule: event sampling; instantaneous sampling; interval recording; rating scales; and duration recording.

Event sampling

Event sampling, also known as a sign system, requires a tally mark to be entered against each statement each time it is observed, for example:

teacher shouts at the child	/////
child shouts at the teacher	///
parent shouts at the teacher	//
teacher shouts at the parent	//

The researcher will need to devise statements that yield the data which answer the research questions. This method is useful for calculating the frequencies or incidence of observed situations or behaviours, so that comparisons can be made; we can tell, for example, that the teacher does most shouting and that the parent shouts least of all. However, whilst these data enable us to chart the incidence of observed situations or behaviours, the difficulty here is that we are unable to

BOX 26.1 NON-PARTICIPANT OBSERVATION: A CHECKLIST OF DESIGN TASKS

1 The preliminary tasks

Have you

- Clearly described the research problem?
- Stated the precise aim of the research?
- Developed an explanation which either links your research to a theory or says why the observations should be made?
- Stated the hypotheses (if any) to be tested?
- Identified the appropriate test statistic (if needed)?

2 The observational system

Have you

- Identified the type(s) of behaviour to be observed?
- Developed clear and objective definitions of each category of behaviour?
- Checked that the categories are complete, and cover all the target behaviours?
- Checked that each category is clearly distinct from the others?
- Checked that the differences between each category are easily seen in the observing situation?

3 The observational process

Have you

- Identified an appropriate location to make your observations?
- Decided which data sampling procedure to use?
- Decided whether to use overt or covert observation?
- Decided whether to use one or more observers to collect information?

4 And finally …

Have you

- Designed the data collection sheet?
- Reviewed the ethical standards of the investigation?
- Run a pilot study and made any necessary amendments to the observation system, or procedure?
- If more than one observer has been used, made a preliminary assessment of inter-observer reliability?

Source: Dyer (1995, p. 186)

determine the *sequence* in which they occurred. For example, two different stories could be told from these data if the sequence of events were known. One story could be seen as follows, where the numbers 1–7 are the different periods over time (e.g. every thirty seconds):

	1	2	3	4	5	6	7
teacher shouts at the child		/	/	/	/		/
child shouts at the teacher	/	/				/	
parent shouts at the teacher		/			/		
teacher shouts at the parent					/	/	

Imagine the scene: a parent and his child arrive late for school one morning and the child slips into the classroom; an event quickly occurs which prompts the child to shout at the teacher, the exasperated teacher is very cross when thus provoked by the child; the teacher shouts at the child who then brings in the parent (who has not yet left the premises); the parent shouts at the teacher for unreasonable behaviour and the teacher shouts back at the child. It seems in this version that the teacher only shouts when provoked by the child or parent.

If the same number of tally marks were distributed in a different order, a very different story might emerge, for example:

	1	2	3	4	5	6	7
teacher shouts at the child	/	/	/	/		/	
child shouts at the teacher					/	/	/
parent shouts at the teacher					/	/	
teacher shouts at the parent			/	/			

In this scene the teacher is the instigator of the shouting, shouting at the child and then at the parent; the child and the parent only shout back when they have been provoked.

Instantaneous sampling

If it is important to know the chronology of events, then it is necessary to use instantaneous sampling, sometimes called time sampling. Here the researcher enters what she observes at standard intervals of time, for example, every twenty seconds, every minute. She notes what is happening at that precise interval moment and enters it into the appropriate category on the schedule. For example, imagine that the sampling will take place every thirty seconds; numbers 1–7 represent each 30-second interval thus:

	1	2	3	4	5	6	7
teacher smiles at the child	/	/	/	/			
child smiles at the teacher			/	/	/	/	
teacher smiles at the parent	/	/	/	/			
parent smiles at the teacher			/	/	/	/	

In this scene the researcher notes down what is happening on the 30-second point and notices from these precise moments that the teacher initiates the smiling but that all parties seem to be doing quite a lot of smiling, with the parent and the child doing the same amount of smiling each. Instantaneous sampling involves recording what is happening on the instant and entering it in the appropriate category. The chronology of events is preserved.

Interval recording

This method charts the chronology of events to some extent and, like instantaneous sampling, requires the data to be entered in the appropriate category at fixed intervals. However, instead of charting what *is happening* on the instant, it charts what *has happened* during the preceding interval. So, for example, if recording were to take place every thirty seconds, then the researcher would note down in the appropriate category what had happened during the preceding thirty seconds. Whilst this enables frequencies to be calculated, simple patterns to be observed and an approximate sequence of events to be noted, because it charts what has taken place in the preceding interval of time, some elements of the chronology might be lost. For example, if three events took place in the preceding thirty seconds of the example, then the order of the three events would be lost; we would know simply that they had occurred.

Wilkinson (2000, p. 236) distinguishes between *whole* interval recording and *partial* interval recording. In the former, behaviour is recorded only if it lasts for the whole of the interval; in the latter, behaviour is recorded if it occupies only a part of the interval in question. In the case of the partial interval recording, the researcher will need to specify how to record this.

Rating scales

Here the researcher makes some judgement about the events being observed, and enters responses into a rating scale. For example, Wragg (1994) suggests that observed teaching behaviour might be entered onto rating scales by placing the observed behaviour onto a continuum:

	1	2	3	4	5	
Warm	___	___	___	___	___	Aloof
Stimulating	___	___	___	___	___	Dull
Businesslike	___	___	___	___	___	Slipshod

An observer might wish to enter a rating according to a five-point scale of observed behaviour, for example:

1 = not at all 2 = very little 3 = a little
4 = a lot 5 = a very great deal

	1	*2*	*3*	*4*	*5*
Child seeks teacher's attention					
Teacher praises the child					
Teacher intervenes to stop misbehaviour					

Here the researcher has to move from low inference (simply reporting observations) to a higher degree of inference (making judgements about events observed). This might introduce a degree of unreliability into the observation (e.g. through: (a) the halo effect; (b) the central tendency, wherein observers will avoid extreme categories; (c) recency, where observers are influenced by more recent events than less recent events). However, this might be a helpful summary way of gathering observational data.

Simpson and Tuson (2003, pp. 42–4) suggest that researchers should ensure that:

- the categories included in ratings adequately cover the 'range of behaviours or features' (p. 42) of interest in the target group for observation;
- the anchor statements (descriptors) on each scale point adequately describe the 'range of possibilities' (p. 43) in the item for observation;
- sufficient specification of what and how to observe is given to researchers for completing the observational schedule, such that two independent observers complete the schedule of the 'same observed activities in the same way' (p. 44).

The duration of behaviour

So far we have concerned ourselves with single events and their recording. This is very suitable for single and usually short-lived behaviours. However, sometimes certain behaviours last a long time and 'over-run' the interval categories or event categories described above, i.e. it is continuous behaviour rather than a single event. For example, a child may remove her shoes only once, but she may continue to be without her shoes for a twenty-minute period; a child may delay starting writing for ten minutes, again a single behaviour but which continues for longer than each of the intervals in interval or instantaneous recording; a child may have a single tantrum which continues for twenty minutes, and so on. What we need is an indication of the *duration* of a particular behaviour. The observation is driven by the event, not the frequency of the observation. This means that the observer needs to structure the recording schedule to indicate the total duration of a single continuous behaviour.

For all the kinds of schedules discussed above, a decision will have to have been agreed in advance on how to enter data. Consistency of entering by a single observer and multiple observers will need to be founded on what counts as evidence, when, where and how to observe, and how many people on whom to focus. Indeed Barrett and Mills (2009) note that having more than one observer is useful for triangulation and reliability. Hill *et al*. (2012) note the importance of reliability checks, as wide variation may be found in data entry on the same observation sheets of the same observed phenomenon by different observers. In turn, this argues for the careful selection and preparation of observers (e.g. Hill *et al*. (2012) note the risk to neutrality in having school principals observing their own teachers).

Further, researchers will need to decide how the observation schedule will distinguish between one person being observed demonstrating the same behaviour twelve times (1 person × 12) or many people demonstrating the same behaviour fewer times (e.g. 2 people × 6 times each, or 4 people × 3 times each), i.e. whether the focus is to be on *people* or on *behaviour*.

Whilst structured observation can provide useful numerical data (e.g. the celebrated ORACLE study) (Galton and Simon, 1980), there are several concerns which must be addressed in this form of observation (Webster, 2015), for example:

- it excludes any mention of the intentions or motivations of the people being observed;
- the individual's subjectivity is lost to an aggregated score;
- it cannot catch certain important features of classrooms and people (e.g. creativity) or important background and contextual factors;
- there is an assumption that the observed behaviour provides evidence of underlying feelings, i.e. that concepts or constructs can be measured in observed occurrences.

This latter point is important, for it goes to the very heart of the notion of validity, since it requires researchers to satisfy themselves that it is valid to infer that a particular behaviour indicates a particular state of mind or particular intention or motivation. As Robson (2014) remarks: observation alone cannot tell us what is in a child's mind (p. 125), and inference can be dangerous. Further, the desire to operationalize concepts and constructs can easily lead researchers to provide overly simple indicators of complex concepts.

Structured observation can also neglect the significance of contexts – temporal and spatial – thereby overlooking the fact that behaviours may be context-specific (Webster, 2015). In their concern for the overt and the observable, researchers may overlook unintended outcomes which may have significance; they may be unable to show how significant are the behaviours of the participants being observed in their own terms, and they may over-simplify and thereby distort behaviour and phenomena (Denscombe, 2014, p. 212). Further, if we accept that behaviour is developmental, that interactions evolve over time and are therefore, by definition, fluid, then the methods of structured observation outlined above appear to take a series of 'freeze-frame' snapshots of behaviour, thereby violating the principle of fluidity of action. Captured for an instant in time, it is difficult to infer a particular meaning to one or more events, just as it is impossible to say with any certainty what is taking place when we study a single photograph or a set of photographs of a particular event. Put simply, the researcher may need to gather additional data from other sources to inform the interpretation of structured observational data. Structured observation runs in tandem with other data-collection methods.

This latter point is a matter not only for structured observation but, equally, for semi-structured and unstructured observation, for what is being suggested here is the point that *triangulation* (of methods, of observers, of time and space) can assist the researcher to generate reliable evidence (Jewitt, 2012; Lee *et al*., 2015). There is a risk that observations may be selective, and the effects of this can be attenuated by triangulation. One way of gathering more reliable data (e.g. about a particular student or group of students) is by *tracking* them through the course of a day or a week, following them from place to place, event to event. Students often behave very differently with one teacher than with another, and a full picture of students' behaviour might require the observer to see the same students in different contexts.

Taking account of criticisms of structured observations, Webster (2015) argues for fitness for purpose: systematic observations are useful in looking at activities that are 'straightforward to identify', 'limited to binary categories' and 'a meaningful expression of behaviour' (p. 995).

26.3 The need to practise structured observation

It may seem to be a naive truism to say that researchers need to practise observation, but they do! For example, they may need to practise entering data in the appropriate categories in the structured observation schedule, and at speed (Simpson and Tuson, 2003, p. 10). They may need to practise where to locate themselves when observing, what to focus on, where to look, what to record (e.g. the level of detail), where to move around the setting, whether to stand or sit, how to code *in situ*, and how to observe without those observed being too conscious of the observation taking place or the observation being too intrusive, what role to take in the classroom or setting, how to avoid eye contact (as this can be threatening or disturbing to the setting or the person being observed), how to observe discreetly or indirectly (e.g. without directly looking intently at a person or group, or without being seen to be looking directly or to be watching or tracking specific persons). Further, observers need to practise and pilot their structured observational instruments in order to find the optimum time intervals for observation schedules for instantaneous sampling, interval recording and duration recording (e.g. five seconds, one minute, two minutes, ten minutes etc.).

26.4 Analysing data from structured observations

For structured observations, researchers can count *frequencies*, for example, with reference to individuals, groups, classes, events, activities, behaviours and so on. One can observe *patterns*, for example in *sequences* of behaviours, conversations or interactions (for instance, by discourse analysis or question-and-answer sequences in classroom talk); or *frequently occurring* combinations of events, behaviours, people, kinds of interaction; or *aggregated data*, for example, from individuals to groups to classes, from individuals to males/females, from individual lessons to courses or subjects, from individual behaviours to categories of behaviour, from individual units of talk to kinds of talk such as closed questions/open questions, extended responses/one-word responses, teacher-initiated talk/student-initiated talk, and on-task/off-task behaviour. Where the data are converted into numbers, the panoply of suitable statistical analyses is available (see Part 5).

In addition to data from structured observations being 'quantitized', they can be turned into narrative accounts, descriptions and themes, i.e. 'qualitized', and we refer readers to Part 5 here.

26.5 Critical incidents

There will be times when reliability-as-consistency in observations is not necessary. For example, a student might only demonstrate a particular behaviour once, but it is so important that it should not be ruled out simply because it occurred once. One only has to commit a single murder to be branded a murderer! Sometimes one event can occur which reveals a very important insight into a person or situation. Critical incidents (Flanagan, 1949; Tripp, 1993) and critical events (Wragg, 1994) are particular events or occurrences that typify or illuminate very starkly a particular feature, for example, a teacher's behaviour or teaching style. Wragg (1994, p. 64) writes that these are events which appear to the observer to have more interest than other ones, and therefore warrant greater detail and recording than other events; they have an important insight to offer. For example, a child might unexpectedly behave very aggressively when asked to work with another child, and this might provide an insight into the child's social tolerance; a teacher might suddenly overreact when a student produces a substandard piece of work – the straw that breaks the camel's back – and this might indicate a level of stress, frustration, tolerance or intolerance, and the effects of that threshold of tolerance being reached. These events are critical in that they may be non-routine but very revealing; they offer the researcher an insight that would not be available by routine observation. They are frequently unusual events or events that have a major impact on what follows them.

26.6 Naturalistic and participant observation

Some observations take place in a context in which the researcher knows clearly and in advance what to look for, with categories and coding worked out before the observation takes place. This is not always the case. It is here that ethnographic and naturalistic observations are pre-eminent (see Chapter 15). Here the intention is to observe participants in their natural settings, their everyday social settings and their everyday behaviour in them. Observation here has a wide embrace, and includes visual observation, document analysis, interviewing, direct observation and introspection (Flick, 2009, p. 226). It is a process, moving from *descriptive observation* (orientation to a field) to *focused observation* (narrowing one's field of observation to focus on those problems and processes that are most germane to the research purpose and questions), and on to *selective observation* (to find further evidence for those items identified in the previous step) (Flick, 2009, p. 227).

Participant observation, as Simpson and Tuson (2003, p. 14) argue, is 'the most subtly intrusive' form of observation since it requires the researcher to be an empathic, sympathetic member of a group, in order to gain access to insiders' behaviours and activities, whilst still acting as a researcher with a degree of detachment. Indeed Watts (2011) suggests that participant observation strives to be non-intrusive, and, since the researcher stays in the situation for a long time, he/she becomes as familiar and unnoticed as everyday objects, part of the furniture, as it were (p. 303). All the participant observer has to do is to be around the place, to listen and watch (p. 303), to be immersed in the locale, to 'hang round' and make field notes (Marshall and Rossman, 2016, p. 143), to take on some participant role in the group – overtly or covertly – and record what was seen, heard, said etc.

Merriam (1998, p. 103) suggests that the participant observer is somewhat 'schizophrenic', as he/she has to balance participation in order to absorb the situation, with sufficient detachment to be able to analyse and observe it in a detached way. It is usually time-consuming, as not only does the researcher have to join in many activities and spend a long time with the group, but he/she has to write up field notes away from the activity itself (e.g. in the evening).

Participant observation is useful for enabling researchers to check their definitions of key terms that are used by participants, to observe events or behaviours that might not be mentioned in interviews, and to gather data on sensitive, unspoken topics (Kawulich, 2005). Participant observation can help in guiding relationships with participants and informants, enable the researcher to 'get a feel' of a situation and how matters are organized in a group or subculture, find out about interactions and relationships, raise questions for further investigation (Schensul *et al*., 1999), sensitize and familiarize a researcher to a context, and reduce reactivity in a short observation (Bernard, 1994). It enables rich descriptions of 'backstage culture' to be gathered (DeMunck and Sobo, 1998, p. 43) and can reveal the dynamics of behaviour, relationships and interactions (Watts, 2011, p. 302; Marshall and Rossman, 2016, p. 143).

As mentioned earlier, there are degrees of participation in observation (LeCompte and Preissle, 1993, pp. 93–4). The 'complete participant' is a researcher

who takes on an insider role in the group being studied, and maybe who does not even declare that she is a researcher (discussed later in comments about the ethics of covert research). The 'participant-as-observer', as its name suggests, is part of the social life of participants, documenting and recording what is happening for research purposes. The 'observer-as-participant', like the participant-as-observer, is known as a researcher to the group, and maybe has less extensive contact with the group. With the 'complete observer', participants may not realize that they are being observed (e.g. using a two-way mirror), hence this is a form of covert research. Hammersley and Atkinson (1983, pp. 93–5) suggest that comparative involvement may come in the forms of the complete participant and the participant-as-observer, with a degree of subjectivity and sympathy, whilst comparative detachment may come in the forms of the observer-as-participant and the complete observer, where objectivity and distance are key characteristics. Both complete participation and complete detachment are as limiting as each other.

As a complete participant, the researcher dare not go outside the confines of the group for fear of revealing her identity (in covert research), whilst as a complete observer there is no contact with the observed, and inference is dangerous. That said, both complete participation and complete detachment minimize reactivity, though in the former there is the risk of 'going native', where the researcher adopts the values, norms and behaviours of the group as her own, i.e. ceases to be a researcher or ceases to be objective (Kawulich, 2005, p. 4), becomes a member of the group and over-identifies with the group (Denscombe, 2014, p. 221).

Participant observation may be particularly useful in studying small groups, or for events and processes that only last a short time or are frequent, for activities that lend themselves to being observed, for researchers who wish to reach inside a situation and have a long time available to them to 'get under the skin' of behaviour or organizations (as in an ethnography), and when the prime interest is in gathering detailed information, both descriptive and explanatory, about what is happening. Participation may be required in order to understand a situation.

In participant observational studies, the researcher stays with the participants for a substantial period of time to reduce reactivity effects (the effects of the researcher on the researched, changing the behaviour of the latter, though Watts (2011) comments that reactivity diminishes quite quickly), and to develop empathy with the culture, values and behaviour of the group under study (Watts, 2011, p. 302). Observers record what is happening, whilst taking a role in that situation. In schools, this might be taking on some particular activities, sharing supervisions, participating in school life, recording impressions, conversations, observations, comments, behaviour, events and activities and the views of all participants in a situation. It is important for the researcher to balance 'participation' with 'observation'; too much of the former compromises the latter, and vice versa (Watts, 2011, p. 303).

Participant observation requires careful attention to gaining access, building trust, identifying a suitable role and being careful about with whom to be seen or with whom to 'hang out' (e.g. the school principal, a marginalized or fringe member of the group) (Kawulich, 2005, pp. 12–13). This latter point extends to the need for care in working with informants, so as not to be at the mercy of informants and gatekeepers. Researchers must recognize that informants may provide selective access to people and to data, and, indeed, depending on the views of the informant by other members of the group, they may prevent access to key people (Flick, 2009, p. 229).

Participant observation is often combined with other forms of data collection that, together, elicit the participants' definitions of the situation and their organizing constructs in accounting for situations and behaviour. By staying in a situation over a long period the researcher can also see how events, behaviours, relationships etc. evolve over time, catching the dynamics of situations, the people, personalities, contexts, resources, roles etc. Morrison (1993, p. 88) argues that by 'being immersed in a particular context over time not only will the salient features of the situation emerge and present themselves but a more holistic view will be gathered of the interrelationships of factors'. Such immersion facilitates the generation of 'thick descriptions' (Geertz, 1973), particularly of social processes and interaction, which lend themselves to accurate explanation and interpretation of events rather than relying on the researcher's own inferences. The data derived from participant observation are 'strong on reality'.

Components of 'thick descriptions' involve recording, for example (Carspecken, 1996, p. 47): speech acts; non-verbal communication; descriptions, using low-inference vocabulary; careful and frequent recording of the time and timing of events; the observer's comments, placed into categories; and detailed contextual data.

Observations, recorded in field notes, can be written at several levels. At the level of *description* they might include, for example (Spradley, 1980; Bogdan and Biklen, 1992, pp. 120–1; LeCompte and Preissle, 1993, pp. 224; Denscombe, 2014; Marshall and Rossman, 2016):

- quick, fragmentary jottings of keywords/symbols;
- transcriptions and more detailed observations written out fully;
- descriptions which, when assembled and written out, form a comprehensive and comprehensible account of what has happened;
- pen portraits of participants;
- reconstructions of conversations;
- descriptions of the physical settings of events;
- descriptions of events, behaviour and activities;
- descriptions of the researcher's activities and behaviour.

Lincoln and Guba (1985, p. 273) suggest a variety of elements or types of observations, including:

- ongoing notes, either verbatim or categorized *in situ*;
- logs or diaries of field experiences (similar to field notes, though usually written some time after the observations have been made);
- notes that are made on specific, predetermined themes (e.g. which have arisen from grounded theory);
- 'chronologs', where each separate behavioural episode is noted, together with the time at which it occurred, or recording an observation at regular time intervals, for example, every two or three minutes;
- context maps – maps, sketches, diagrams or some graphic display of the context (usually physical) within which the observation takes place, such graphics enabling movements to be charted;
- entries on predetermined schedules (including rating scales, checklists and structured observation charts), using taxonomic or categoric systems, where the categories derive from previous observational or interview data;
- sociometric diagrams (indicating social relationships, e.g. isolates (whom nobody chooses), stars (whom everyone chooses) and dyads (who choose each other));
- debriefing questionnaires from respondents that are devised for, and by, the observer only, to be used for reminding the observer of the main types of information and events once she or he has left the scene;
- data from debriefing sessions with other researchers, as an aide-memoire.

LeCompte and Preissle (1993, pp. 199–200) provide useful guidelines for directing observations of specific activities, events or scenes, suggesting that they should include answers to the following questions:

- Who is in the group/scene/activity, who is taking part?
- How many people are there, their identities and their characteristics?
- How do participants come to be members of the group/event/activity?
- What is taking place?
- How routine, regular, patterned, irregular and repetitive are the behaviours observed?
- What resources are being used in the scene?
- How are activities being described, justified, explained, organized, labelled?
- How do different participants behave towards each other?
- What are the statuses and roles of the participants?
- Who is making decisions, and for whom?
- What is being said, and by whom?
- What is being discussed frequently/infrequently?
- What appears to be the significant issues that are being discussed?
- What non-verbal communication is taking place?
- Who is talking and who is listening?
- Where does the event take place?
- When does the event take place?
- How long does the event take?
- How is time used in the event?
- How are the individual elements of the event connected?
- How are change and stability managed?
- What rules govern the social organization of, and behaviour in, the event?
- Why is this event occurring, and occurring in the way that it is?
- What meanings are participants attributing to what is happening?
- What are the history, goals and values of the group in question?

That this list is long (and by no means exhaustive) reflects the complexity of even the most apparently mundane activity. It can sensitize observers.

Lofland (1971) suggests six main categories of information in participant observation:

- acts (specific actions);
- activities (which last a longer time, for instance, a week, a term, months, e.g. attendance at school, membership of a club);
- meanings (e.g. how participants explain the causes, meanings and purposes of particular events and actions);
- participation (what the participants do, e.g. membership of a family group, school groups, peer group, clubs and societies, extra-curricular groups);

- relationships (those which are observed in the several settings and contexts in which the observation is undertaken);
- settings (descriptions of the settings of the actions and behaviours observed).

Spradley (1980, p. 78) suggests a checklist of contents of field notes:

- space: the physical setting;
- actors: the people in the situation;
- activities: the sets of related acts that are taking place;
- objects: the artefacts and physical things that are there;
- acts: the specific actions that participants are doing;
- events: the sets of activities that are taking place;
- time: the sequence of acts, activities and events;
- goals: what people are trying to achieve;
- feelings: what people feel and how they express this.

The context of observation is important (Silverman, 1993, p. 146). Moyles (2002, p. 181) suggests that researchers should record the physical and contextual setting of the observation, the participants (e.g. number, who they are, who comes and goes, what they do and what are their roles), the time of day of the observation, the layout of the setting (e.g. seating arrangements, arrangement of desks), the chronology of the events observed, and any critical incidents that happened.

At the level of *reflection*, field notes can include (Bogdan and Biklen, 1992, p. 122):

- reflections on the descriptions and analyses that have been done;
- reflections on the methods used in the observations and data collection and analysis;
- ethical issues, tensions, problems and dilemmas;
- reactions of the observer to what has been observed and recorded – attitude, emotion, analysis etc.;
- points of clarification that have been and/or need to be made;
- possible lines of further inquiry.

Lincoln and Guba (1985, p. 327) indicate three main types of item that might be included in a journal:

- a daily schedule, including practical matters, for example, logistics;
- a personal diary, for reflection, speculation and catharsis;
- notes on, and a log of, methodology.

In deciding what to focus on, Wilkinson (2000, p. 228) suggests an important distinction between observing *molecular* and *molar* units of behaviour. Small units of behaviour are molecular, for example, gestures, non-verbal behaviour, short actions, short phrases of a conversation. Whilst these yield very specific data, they risk being taken out of context, such that their meanings, and thereby their validity, are reduced. By contrast, the molar approach deals in large units of behaviour, the size of which is determined by the theoretical interests of the researcher. The researcher must ensure that the units of focus are valid indicators of the issues of concern to the researcher.

From all this we suggest that observational data should be comprehensive enough to enable the reader to reproduce the analysis that was performed. It should focus on the observable and make explicit the inferential, and the construction of abstractions and generalizations might commence early but should not starve the researcher of novel channels of inquiry (Sacks, 1992).

Spradley (1979) and Kirk and Miller (1986) suggest that observers should keep four sets of observational data to include: notes made *in situ*; expanded notes that are made as soon as possible after the initial observations; journal notes to record issues, ideas, difficulties etc. that arise during the fieldwork; and a developing, tentative running record of ongoing analysis and interpretation.

The intention here is to introduce some systematization into observations in order to increase their reliability. In this respect, Silverman (1993) reminds us of the important distinction between *etic* and *emic* analysis. *Etic* analysis uses an objective conceptual framework, perhaps that of the researcher, whilst *emic* approaches use the subjective conceptual frameworks of those being researched. Structured observation uses *etic* approaches, with predefined frameworks that are adhered to unswervingly, whilst *emic* approaches sit comfortably within qualitative approaches, where the definitions of the situations are captured through the eyes of the observed.

Participant observation is not without its critics, being described as subjective, biased, impressionistic, idiosyncratic and lacking in the precise quantifiable measures that are the hallmark of survey research and experimentation. Whilst it is probably true that nothing can give better insight into the life of a gang of juvenile delinquents than going to live with them for an extended period of time, critics of participant observation studies will point to the dangers of 'going native' as a result of playing a role within such a group. How do we know that observers do not lose their perspective and become blind to the peculiarities that they are

supposed to be investigating? How do they 'fight familiarity'?

Further, Johnson and Sackett (1998) suggest that participant observation risks being selective, unrepresentative and more concerned with the agenda of the researcher rather than the real situation (they report that researchers were more concerned with commenting on political and religious behaviours than with eating and sleeping behaviours, yet political and religious behaviours accounted for only 3 per cent of the participants' time whilst eating and sleeping accounted for 60 per cent of their time).

Adler and Adler (1994, p. 380) suggest several stages in an observation. Commencing with the selection of a setting on which to focus, the observer then seeks a means of gaining entry to the situation (e.g. taking on a role in it). Having gained entry the observer can then commence the observation proper, be it structured or unstructured, focused or unfocused. If quantitative observation is being used then data gathered can be analysed *post hoc*; if more ethnographic techniques are being used then *progressive focusing* requires the observer to undertake analysis *during* the period of observation itself.

The question that researchers frequently ask is 'how much observation must I do?' or 'when do I stop observation?'. Of course, there is no hard and fast rule here, though it may be appropriate to stop when 'theoretical saturation' has been reached (Adler and Adler, 1994, p. 380), i.e. when the situations that are being observed appear to be repeating data or issues that have already been collected (see also Chapter 37). It may be important to carry on collecting data at this point, to indicate overall frequencies of observed behaviour, enabling the researcher to find the most to the least common behaviours observed over time. Further, the greater the number of observations, the greater the reliability of the data might be, enabling emergent categories to be verified. What is being addressed here is the reliability of the observations (see Chapter 14 on triangulation).

26.7 Data analysis for unstructured observations and videos

For less structured and unstructured observational data (e.g. from field notes, videos), the tools of qualitative analysis can be used, for example: summarizing; narrative accounts (of individuals, groups, behaviours, events); thematic analysis and patterning; coding and categorizing; nodes and connections; constant comparison; theoretical saturation (see Part 5). This includes use of computer-based software for analysing qualitative processing (e.g. NVivo, ATLAS.ti, MAXQDA) (see also: www.surrey.ac.uk/sociology/research/researchcentres/caqdas/support/choosing/index.htm). For video material, NVivo, Orion, Transana and VideoTrace are also available at the time of writing.

Simpson and Tuson (2003, pp. 83–5) and Miles and Huberman (1984) indicate several strategies for data analysis of field notes and qualitative data, including, largely with reference to coding:

- reviewing, analysing and coding early rather than accumulating too much data before analysis;
- coding densely at first (i.e. avoiding moving too quickly into summarizing);
- keeping track of the data analysis over time (e.g. key codes and what they embrace, key people observed, keeping to the research questions (if appropriate, i.e. depending on the nature of the research));
- verifying intuitions with data;
- identifying themes and patterns (sometimes by counting frequencies or consistencies);
- looking for clusters of events, activities, people, behaviours etc.;
- writing metaphors to catch the essence of features;
- being prepared to disaggregate as well as aggregate data in order to preserve fidelity to the events/people/situations;
- putting codes into hierarchies (some codes are subsumed by others);
- ensuring conceptual coherence to the analysis.

Merriam (1998) suggests that it is useful for researchers to identify keywords, not only in the observed events, but in their analysis, together with attention to the start and end of observed conversations, as these are often significant and most easily remembered. Kawulich (2005) reports the value of 'quantitizing' data, looking for frequencies, together with narrative descriptions of settings, participants, activities and behaviours. She commends the use of two types of field notes for analysis: (a) observed data, including verbatim conversations; and (b) reflections, questions to be asked, issues for further exploration and comments (i.e. observations on observations). Hence observational data can be both mixed methods in themselves and in conjunction with other methods of data collection and analysis.

26.8 Natural and artificial settings for observation

Most observations by educational researchers will be undertaken in natural settings: schools, classrooms, playgrounds, lessons and suchlike. In studies of a psychological nature it may be that a contrived, artificial,

perhaps laboratory setting is organized in order to give greater observational power to the observers.

Some observational behavioural research uses a two-way mirror, in which those being observed see a mirror on a wall through which unseen observers watch what is happening without disturbing the setting under observation (e.g. counsellor training, young children interacting with each other, parents with their children), thereby causing anxiety among the participants. Often rooms are specifically prepared for this, and they may also include video camera installations. It raises questions of the ethics of covert research, discussed below.

In Chapter 30 we describe two classic studies in the field of social psychology, both of which used contrived settings: the Milgram study of obedience to authority and the Stanford Prison Experiment. Similarly, psychological researchers may wish to construct a classroom with a two-way mirror in order to observe children's behaviour without the presence of the observer. Again, this raises the ethical issue of overt and covert research. The advantage of a contrived, artificial setting is the degree of control that the researcher can exert over the situation, often as much as in a laboratory experiment. To the charge that this is an unrealistic situation and that humans should neither be controlled nor manipulated, we refer the reader to the ethical issues addressed in Chapter 7.

Settings may be classified by the degree of structure that is imposed on the environment by the observer/researcher, and by the degree of structure inherent in the environment itself (Cooper and Schindler, 2001, p. 378). Table 26.2 places settings for observation along a continuum from structured to unstructured and from natural to artificial.

Clearly the researcher will need to be guided by 'fitness for purpose' in the type of setting and the amount of structure imposed. There is fuzziness between the boundaries here. As mentioned earlier, structured settings may be useful in testing hypotheses whilst unstructured settings may be useful for generating hypotheses.

26.9 Video observations

In addition to the observer writing down details in field notes, audio-visual recording is useful (Erickson, 1992, pp. 209–10). Video recording equipment is now relatively cheap and accessible, and it provides a 'fine-grained multimodal record of an event detailing gaze, expression, body posture, and gesture' (Jewitt, 2012, p. 6).

Video recording can overcome the partialness of the observer's view of a single event (a video can be shared by several researchers) and can overcome the tendency towards only recording the frequently occurring events. Video recording can offer a more 'unfiltered' observational record of 'natural' human behaviour in real time, and it maintains the sequence of the event (Simpson and Tuson, 2003, p. 51; Jewitt, 2012; Blikstad-Balas, 2016). The video record can be viewed several times; it is not a 'once-and-for-all' observation. Video data have the capacity for completeness of analysis and comprehensiveness of material, reducing the dependence on prior interpretations by the researcher and enabling the researcher to scrutinize data.

On the other hand, one has to be cautious here, for installing video cameras might create the problem of reactivity (e.g. Jewitt, 2012; Lee *et al*., 2015); even if it is not obvious to the observer and even if people are not looking at the camera, their behaviour might change if they are being videoed, for example, they may behave in a socially desirable or deliberately acceptable way. Further, if fixed cameras are used, they might be as selective as participant observers, including and excluding areas of focus, even if the cameras are moveable (Jewitt, 2012). Like other forms of observation,

TABLE 26.2 STRUCTURED, UNSTRUCTURED, NATURAL AND ARTIFICIAL SETTINGS FOR OBSERVATION

	Natural setting	*Artificial setting*
Structured	Structured field studies (e.g. Sears *et al.*'s (1965) study of *Identification and Child Rearing*)	Completely structured laboratory (e.g. the Stanford Prison experiment, the Milgram experiment on obedience, see Chapter 26). Experiments with one-way mirrors or video recordings
Unstructured	Completely unstructured field study (e.g. Whyte's (1993) celebrated study of *Street Corner Society*, and ethnographic studies)	Unstructured laboratory (e.g. Axline's (1964) celebrated study of *Dibs in Search of Self*. Observations with one-way mirrors or video recordings)

the video only records what is observable – the material world – and this is only part of a situation; it 'frames' an event (p. 8), and this may risk losing the context in which the event is located.

Whilst a human observer can turn his/her attention to an event that occurs, for example, in a different part of the classroom, a fixed video camera cannot, and indeed a movable camera that changes direction to focus on that event or group of students might be very intrusive. Further, students, unintentionally, might block the camera's eye or move across the classroom and 'get in the way' of the focus of the camera, such that the observation is lost, whereas a human can see much more easily. Whilst having a second or third camera in the classroom might overcome this, it may be costly in terms of equipment, time needed to review and analyse the recordings, and intrusiveness.

Video cameras can be set in close-up focus to catch certain details (e.g. facial expressions), but this rules out the benefits of a panoramic focus (e.g. to catch other class members or activities); on the other hand, a panoramic focus may not have the degree of focus required for close-up detail.

Also available to some researchers is surveillance video footage (e.g. CCTV) data, with access, suitable permission, clearance, ethical approval and secure technical playback facilities. For example, schools are increasingly installing CCTV facilities; this unobtrusive measure is cost-efficient, catches hitherto 'blind spots' in locations, does not require a camera operator to be present and, because it is often archived, can also enable longitudinal studies to be conducted (though many CCTV systems will only store video data for a fixed period of time before being deleted) (Lee *et al*., 2015). The quality of such video images may also be poor, and the absence of audio material renders the video record incomplete.

To overcome the limitations of fixed and moveable cameras, another option is to have wearable cameras – a point of view (POV) camera – for example, students can wear headgear which has a mini-camera, or they may have digital pens which hold a small camera (Maltese *et al*., 2016). This enables the gaze of the camera to follow the focus, attention and direction of the wearer's gaze (Lahlou, 2011; Maltese *et al*., 2016). However, this overt form of observation may affect the behaviour of the participants even more than a fixed camera: the reactivity issue (though Heath *et al*. (2010) suggest that reactivity is exaggerated and reduces quickly).

Jewitt (2012), Robson (2014) and Lee *et al*. (2015) suggest that researchers use video observation in conjunction with other data sources in order to gather a rounded and triangulated picture of a situation, particularly if the video has no soundtrack. Video recordings, whilst enabling thick descriptions to be obtained, take time to watch and analyse, are heavy on detail and high granularity (risking data overload), and typically the video recording itself might comprise short rather than long periods of time, i.e. videos are selective in time events; hence they may overemphasize small details and lose the 'big picture' which is only obtained by longer-term observation (Lemke, 2007), though some of this challenge can be overcome by time-lapse videoing (Jewitt, 2012, p. 5). This also raises the issue of when to start and stop the video recording (Flick, 2009, p. 251).

In interpreting video material, the researcher may need expertise and suitable experience to make sense of the material. For example, Schindler (2009) reports the case of researchers who missed and did not understand data, in this instance concerning small physical movements performed in martial arts, because they were insufficiently knowledgeable about martial arts. We address video analysis in Part 5.

Researchers using video observation will need to decide (cf. Derry *et al*., 2010; Jewitt, 2012; Lee *et al*., 2015; Blikstad-Balas, 2016):

- what kind of recording to use (visual only, audio-visual, overt or covert, CCTV etc.);
- the focus of the video (e.g. close-up, distant) and how to balance close-up and long-distance focus;
- how many cameras, and what kind (e.g. fixed, moveable, wearable, digital pens etc.);
- whether to have a fixed, roaming or wearable camera;
- who operates the video camera(s);
- where to position the camera(s);
- when to start and stop the recording;
- how many events to record and over what time period (i.e. how much data to collect and from whom);
- how to catch the context of the video recording and the 'bigger picture' over time;
- how to avoid data overload;
- how to minimize reactivity;
- how to combine video data with other data to obtain a complete picture;
- the unit of analysis for the video, for example, individuals groups, events, behaviours, time, themes, etc.;
- how to analyse, interpret and report the video data.

26.10 Timing and causality with observational data

Observation in experimental procedures is prone to problems of timing: too soon and the effect may not be noticed or may be too short-term; too late and the effect might have gone or been submerged by other matters. Experiments typically suffer from the problem of only having two time points for observational measurement: the pre-test and the post-test, and this offers researchers little opportunity for identifying causal *processes* and *mechanisms* at work. The choice of timing of observation for establishing causation is crucial and it varies with the purposes of the research. The frequency of the observational data collection varies with the phenomenon under investigation, the scope of the phenomenon, the overall timescale of the phenomenon, the speed at which the dependent variable is likely to change and the level of detailed causal explanation required.

Sometimes micro-time is important (e.g. intervals of just a few seconds, as in the data collection for the ORACLE studies) (Galton and Simon, 1980). In other research a longer time frame is more suitable. Rather than fixing a specific time, it may be the events themselves that dictate the timing of the data collection, so that, for example, changes are reported when they occur, which may vary in time.

The first rule of thumb here is that the more accurately we wish to know the causal sequences, the more frequently and closer together must be the observational data-collection points. As the number of time points for data collection increases, so does the likelihood of making correct causal inferences and establishing correct causal processes and causation (Morrison, 2009, p. 168).

The second rule of thumb is that the more complex the phenomenon under investigation is, i.e. the more possible causal lines there are in a network of causation (Morrison, 2012), the more time points for observational data collection might be necessary in order to understand the causation at work. Hage and Meeker (1988, p. 177) comment that most causal processes are either not observable or not easily observable, i.e. inference overrules description. The shorter and more frequent are the time intervals and times of data collection respectively, the more the causal inferences become a matter of informed inference rather than of faith.

If we wish to understand causation at work then rich data are necessary. Hence, concomitant with the first two rules of thumb is the third rule of thumb: the more we wish to understand causation and causal processes, then the more it is that qualitative observational data can be useful, as they often have much greater explanatory potential than numerical data. Qualitative observational data can be ongoing and in-depth, and they can indicate causation at work, action narratives and agency within broader conditions and constraints. Consider clinical case studies of individuals, which may have masses of rich qualitative, observational data and field notes that thereby enable researchers to understand the processes and mechanisms of causation at work. Participant observation, rather than being peripheral in the battery of data-collection methods, becomes important in understanding causation at work. This is potentiated when used in combination with other qualitative methods (e.g. Hage and Meeker 1988, p. 179), not least because observation on its own does not establish causation, as much causation is unobservable.

Ethnography may have the edge over experimentation in understanding causal processes in the real world of education rather than the laboratory. The case for qualitative observational data in understanding causation and causal processes is powerful, even pre-eminent.

26.11 Ethical considerations in observations

There are several ethical considerations surrounding observation, and, typically, ethics committees for research will need to give clearance for the observation(s) to happen (Pearson, 2009; Derry *et al*., 2010; Jewitt, 2012; Marshall and Rossman, 2016). To undertake observation, as with many other forms of data collection, requires the informed consent of participants, the right not to be observed, permission from the school and the parents, and perhaps clearance concerning the researcher's reliability and safety to work with young people in schools. All of these take on even greater significance if the researcher is to conduct participant observation or if the research involves close-up observation, for example, observation which might invade the personal space of participants or contain any sense of threat (Simpson and Tuson, 2003, p. 61). Informed consent also has to attend to the cultural dimension of observation, for example knowing whom to approach, how to address them, how to secure permission in a culturally appropriate manner, and so on. However, it may simply be impractical to gain informed consent (e.g. if there are large groups of people being observed, such as in a public place, or if surveillance video data are being used (Watts, 2011, p. 305) or if covert research is being undertaken).

There is much literature on the dilemma surrounding overt and covert observation. Whereas in overt research the subjects know that they are being observed,

in covert research they do not. On the one hand, covert research appears to violate the principle of informed consent, invades the privacy of subjects and private space, treats the participants instrumentally – as research objects – and places the researcher in a position of misrepresenting her/his role (Mitchell, 1993), or rather, of denying it. On the other hand, there are some forms of knowledge that are legitimately in the public domain but access to which is only available to the covert researcher, for example, the fascinating account of the lookout 'watch queen' in the homosexual community (Humphreys, 1975). Covert research might be necessary to gain access to marginalized and stigmatized groups, or groups who would not willingly accede to the requests of a researcher to become involved in research. This might include those groups in sensitive positions, for example drug users and suppliers, HIV sufferers, political activists, child abusers, police informants and human traffickers.

Mitchell (1993) makes a powerful case for covert observational research, arguing that not to undertake covert research is to deny access to groups who operate under the protection of silence, to neglect research on sensitive but important topics and to reduce research to mealy-mouthed avoidance of difficult but strongly held issues and beliefs, i.e. to capitulate when the going gets rough. In a series of examples of covert research, Mitchell makes the case that not to undertake this kind of research is to deny the public access to areas of legitimate concern, the agendas of the powerful (who can manipulate silence and denial of access to their advantage) and public knowledge of poorly understood groups or situations.

Covert observation can also be justified on the grounds that it overcomes problems of reactivity, particularly if the researcher believes that individuals would change their natural behaviour if they knew that they were being observed. In some cases covert observation can produce more reliable, less biased results than overt observation, and indeed may be justified where the safety of the researcher may be at risk (Pearson, 2009).

That covert research can be threatening is well documented. For example, Patrick's (1973) study of a Glasgow gang, where the researcher had to take great care not to 'blow his cover' when witness to a murder, or Mitchell's (1993) research on mountaineers, where membership of the group involved initiation into the rigours and pains of mountaineering (the researcher had to become a fully fledged mountaineer to gain acceptance by the group).

Ethical issues also have to address the problem that observations often disturb the natural setting. Bernard (1994) suggests that participation may involve some deception, pretence and impression management in order to achieve rapport, access, immersion, objectivity and an ability to blend into the community or context, even if the researcher is overt rather than covert. The observer may have to feign ignorance or willingness to be involved in order to gain access to sensitive or confidential data (thereby using persons as objects rather than as subjects).

The ethical dilemmas of covert research are numerous, charting the tension between invasion and protection of privacy and the public's legitimate 'right to know', between informed consent and breaking this in the interests of a wider public. At issue is the dilemma that arises between protecting the individual and protecting the wider public, posing the question 'whose beneficence?': whom does the research serve or protect; is the greater good the protection and interests of the individual or the protection and interests of the wider public; will the research harm already damaged or vulnerable people or will it improve their lot; will the research have to treat people instrumentally in the interests of gathering otherwise unobtainable yet valuable research data? Should the researcher disclose all the data (Kawulich, 2005) or keep some private? (Kawulich (2005, p. 14) reports being told that she should not request additional funding for research if the research was not publishable, and she decided not to publish some data in order to retain good relationships with the group she was studying; her loyalty was to the group rather than to the public.) The researcher has inescapable moral obligations to consider, and, whilst ethical codes abound, each case must be judged on its own merits. Issues of disclosure concern confidentiality, non-traceability, protection of identity and, principally, the ethic of *primum non nocere*: first do no harm.

The need for covert research, with due protections, is justified ethically in guidelines and codes of ethics. For example, the British Sociological Association (2002, para. 31) indicates that 'the use of covert methods may be justified in certain circumstances' and that 'covert methods violate the principles of informed consent and may invade the privacy of those being studied. Covert researchers might need to take into account the emerging legal frameworks surrounding the right to privacy' (para. 32). The British Educational Research Association (2011) writes that 'researchers must therefore avoid deception or subterfuge unless their research design specifically requires it to ensure that the appropriate data is collected or that the welfare of the researcher is not put in jeopardy' (para. 14). The American Educational Research Association's *Ethical*

Standards (2000) state: 'Deception is discouraged; it should be used only when clearly necessary for scientific studies, and should then be minimized. After the study, the researcher should explain to the participants and institutional representatives the reasons for the deception' (para. 3); here the requirement for full subsequent disclosure might prevent certain kinds of research from being done. Pearson (2009, p. 244) comments that, in considering covert research, 'proportionality' has to be addressed, whereby the potential harm done to individuals and organizations is minimal, and much less than the public benefit to be gained from the research.

The issue of non-intervention is also ethically problematic. Whilst the claim for observation as being non-interventionist was made at the start of this chapter, the issue is not as clean as this, for researchers inhabit the world that they are researching, and their influence may not be neutral (the Hawthorne and halo effects discussed in Chapter 14). This is clearly an issue in, for example, school inspections, where the presence of an inspector in the classroom exerts a powerful influence on what takes place; it is disingenuous to pretend otherwise. Observer effects can be considerable.

Moreover, the non-interventionist observer has to consider carefully her/his position. In the example of Patrick's witness to a murder mentioned above, should the researcher have 'blown his cover' and reported the murder? What if not acting on the witnessed murder might have yielded access to further sensitive data? Should a researcher investigating drug or child abuse report the first incident or 'hang back' in order to gain access to further, more sensitive data? Should a witness to abuse simply report it or take action about it? If I see an incident of bullying, do I maintain my non-interventionist position? Do I 'turn a blind eye' to breaches of discipline or school rules (e.g. an individual's or group's plans to bully a student, to physically assault someone, to steal and so on), or even criminal acts or plans for criminal acts (Pearson, 2009, p. 246)? Do I undertake criminal acts in order to be an insider to a group (Pearson, 2009)? Is the observer merely a journalist, providing data for others to judge? When does non-intervention become morally reprehensible? Just because observation may not be illegal (e.g. photographing a couple kissing intimately in a public place), does this make it acceptable? These are issues for which one cannot turn to codes of conduct for a clear adjudication.

26.12 Reliability and validity in observations

Many observation situations carry the risk of bias (e.g. Wilkinson, 2000, p. 228; Moyles, 2002, p. 179; Robson, 2002, pp. 324–5; Shaughnessy *et al*., 2003, pp. 116–17; Flick, 2009, chapter 17; Jewitt, 2012; Breznau, 2016), for example by:

- *selective attention* of the observer: what we see is a function of where we look, what we look at, how we look, when we look, what we think we see, whom we look at, what is in our minds at the time of observation; what are our own interests and experiences;
- *reactivity*: participants may change their behaviour if they know that they are being observed, for example, they may try harder in class, they may feel more anxious, they may behave much better or much worse than normal, they may behave the way they think the researcher wishes them to, or in ways for which the researcher tacitly signals approval (Shaughnessy *et al*., 2003, p. 113);
- *attention deficit*: what if the observer is distracted, or looks away and misses an event?
- *validity of constructs*: decisions must be taken on what counts as valid evidence for a judgement. For example, is a smile a relaxed smile, a nervous smile, a friendly smile, a hostile smile? Does looking at a person's non-verbal gestures count as a valid indicator of interaction? Are the labels and indicators used to describe the behaviour of interest valid indicators of that behaviour?
- *selective data entry*: what we record can be affected by our personal judgement rather than the phenomenon itself; we may interpret the situation and then record our interpretation rather than the phenomenon;
- *selective memory*: if we write up our observations after the event, our memory neglects and selects data, sometimes overlooking the need to record the contextual details of the observation; notes should be written either during or immediately after the observation;
- *interpersonal matters and counter-transference*: our interpretations are affected by our judgements and preferences – what we like and what we don't like about people and their behaviour – together with the relationships that we may have developed with those being observed and the context of the situation; researchers may have to deliberately distance themselves from the situation and address reflexivity;
- *expectancy effects*: the observer knows the hypotheses to be tested, or the findings of similar studies, or has expectations of finding certain behaviours, and

these may influence her/his observations. Expectancy effects can be overcome by ensuring that the observers do not know the purpose of the research, the 'double-blind' approach;

- *decisions on how to record*: the same person in a group under observation may be demonstrating the behaviour repeatedly, but nobody else in the group may be demonstrating that behaviour; there is a need to record how many different people show the behaviour;
- *number of observers*: different observers of the same situation may be looking in different directions, so there may be inconsistency in the results. Therefore there is a need for training, for consistency, for clear definition of what constitutes the behaviour, of entry/judgement and for kinds of recording;
- *the problem of inference*: observations can only record what happens and what can be seen, and it may be dangerous, without any other evidence, for example, triangulation, to infer reasons, intentions, causes and purposes that lie behind actors' behaviours. One cannot always judge intention from observation, for example, a child may intend to be friendly, but it may be construed by an inexperienced observer as selfishness; a teacher may wish to be helpful but the researcher may interpret it as threatening. It is dangerous to infer a stimulus from a response, an intention from an observation. Similarly, one may not see certain phenomena emerging over time (e.g. biographical processes);
- *difference of interpretation of, and data aggregation and conclusions from, the same data*: Breznau (2016, p. 302) terms these 'secondary observer effects'.

The issues here concern validity and reliability. With regard to the validity of the observation, researchers have to ensure that the indicators of the construct under investigation are fair and operationalized, for example, there is agreement on what counts as constituting qualities such as 'friendly', 'happy', 'aggressive', 'sociable' and 'unapproachable'. The matter of what to observe is problematic. For example, do you only focus on certain people rather than the whole group, on certain events and at certain times rather than others, on molar or molecular units? Do you provide a close-grained, close-up observation or a holistic, wider-focused and wider-ranging observation, i.e. do you use a zoom lens and obtain high definition of a limited scope, or a wide-angle lens and obtain a full field but lacking in detail, or somewhere between the two? How do you decide on what to focus?

With regard to reliability, the indicators have to be applied fully, consistently and securely, with no variation in interpretation. Reliability resides not only in the instrument but their use by different raters (Hill *et al*., 2012). This is a matter not only for one observer – consistency in his or her observation and recording – but also if there are several observers. A formula for calculating the degree of agreement (as a percentage) between observers can be used, thus:

$$\frac{\text{Number of times two observers agree}}{\text{Number of possible opportunities to agree}} \times 100$$

In measuring inter-rater reliability one should strive for a high percentage (over 90 per cent minimum). Other measures of inter-rater reliability use correlations, and here coefficients of >0.90 (i.e. over 90 per cent) should be sought (Shaughnessy *et al*., 2003, p. 111). Hill *et al*. (2012) remind researchers that rater agreement measurement does not necessarily provide a complete picture of inter-rater reliability as there are other variables included (e.g. frequency of observations, frequency of behaviours observed, cognitive load on raters, number of items, random variation etc.).

To ensure reliability, it is likely that training is required, so that researchers:

- use the same operational definitions;
- record the same observations in the same way;
- look for the same things;
- have good concentration;
- can focus on detail;
- can be unobtrusive but attentive;
- have the necessary experience to make informed judgements from the observational data.

These qualities are essential in order to avoid fatigue, 'observer drift' (Cooper and Schindler, 2001, p. 380) and halo effects, all of which can reduce reliability.

With regard to reactivity, one suggestion is to adopt covert observation, though this raises ethical issues (see Chapter 7 and above). Another suggestion is to adopt habituation, i.e. the researcher remains in the situation for such a long time that participants become used to his/her presence and revert to their natural behaviour.

To aid reliability in the research, it is also important for the observer to write up notes as soon after the event as possible (writing may stimulate more thought), to write quickly yet to expect to take a long time to write notes, to use computer software (for subsequent data processing) and to make two copies: one of the original data and another for manipulation and analysis (e.g. cutting and pasting data).

26.13 Conclusion

Observation is a powerful tool for gaining insight into situations. As with other data-collection techniques, observation engages issues of validity and reliability. Even low inference observation, perhaps the safest form of observation, is itself selective, just as perception is selective. Higher forms of inference, whilst moving towards establishing causality, rely on greater levels of interpretation by the observer, with the observer making judgements about intentionality and motivation. In this respect, additional methods of gathering data can be employed, to provide corroboration and triangulation, in short, to ensure that reliable inferences are derived from reliable data.

In planning observations one has to consider:

- when, where, how and what to observe;
- how much degree of structure is necessary in the observation;
- the duration of the observation period, which must be suitable for the behaviour to occur and be observed;
- the timing of the observation period (e.g. morning, afternoon, evening);
- the context of the observation (a meeting, a lesson, a development workshop, a senior management briefing etc.);
- the nature of the observation (structured, semi-structured, open, molar, molecular etc.);
- the need for there to be an opportunity to observe, for example, to ensure that there is the presence of the people/behaviour to be observed;
- the merging of subjective and objective observation, even in a structured observation: an observation schedule can become highly subjective when it is being completed, as interpretation, selection and counter-transference may enter the observation, and operational definitions may not always be sufficiently clear;
- the value of covert participant observation in order to gain access and to reduce reactivity;
- threats to reliability and validity;
- the need to operationalize the observation so that what counts as evidence is consistent, unambiguous and valid, for example, what constitutes a particular quality (e.g. anti-social behaviour: what counts as anti-social behaviour – one person's 'sociable' is another's 'unsociable' and vice versa);
- the need to choose the appropriate kind of structured observation and recording (e.g. event sampling, instantaneous sampling, whole interval/partial interval recording, duration recording, dichotomous/rating scale recording);
- how to go under cover, or whether informed consent is necessary;
- ethically defensible observation;
- whether deception is justified;
- which role(s) to adopt on the continuum of complete participant, to participant-as-observer, to observer-as-participant, to complete observer.

Observation can be a very useful research tool. On the other hand, it exacts its price, for example: it may take a long time to catch the required behaviour or phenomenon, it can be costly in time and effort and prone to difficulties of interpreting or inferring what the data mean. This chapter has outlined several different types of observation and the premises that underlie them, the selection of the method to be used depending on 'fitness for purpose'. Overriding the issues of which specific method of observation to use, this chapter has suggested that observation places the observer into the moral domain, that it is insufficient simply to describe observation as a non-intrusive, non-interventionist technique and thereby to abrogate responsibility for the participants involved. Like other forms of data collection in the human sciences, observation is not a morally neutral enterprise. Observers, like other researchers, have obligations to participants as well as to the research community.

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Tests

CHAPTER 27

The field of testing is so extensive that the comments that follow must needs be of an introductory nature; the reader seeking a deeper understanding will need to refer to specialist texts and sources on the subject. Limitations of space permit no more than a brief outline of some key issues concerning tests and testing, including:

- what are we testing?
- parametric and non-parametric tests
- diagnostic tests
- norm-referenced, criterion-referenced and domain-referenced tests
- commercially produced tests and researcher-produced tests
- constructing and validating a test
- software for preparation of a test
- devising a pre-test and post-test
- ethical issues in testing
- computerized adaptive testing

27.1 Introduction

Since the spelling test of Rice (1897), the fatigue test of Ebbinghaus (1897) and the intelligence scale of Binet (1905), the growth of tests has proceeded at an extraordinary pace in terms of volume, variety, scope and sophistication. In tests, researchers have at their disposal a powerful method of data collection (and indeed secondary data) and an impressive array of tests for gathering data of a numerical rather than verbal kind. In considering testing for gathering research data, several issues need to be borne in mind, not least of which is why tests are being used at all:

- What are we testing (e.g. achievement, aptitude, attitude, personality, performance, intelligence, social adjustment etc.)?
- Are we dealing with parametric or non-parametric tests?
- Are they norm-referenced or criterion-referenced?
- Are they available commercially for researchers to use or will researchers have to develop home-produced tests?
- Do the test scores derive from a pre-test and post-test in the experimental method?
- Are they group or individual tests?
- Do they involve self-reporting or are they administered tests?
- How to construct and validate a test?

We unpack some of these issues in this chapter.

Larger-scale testing (e.g. national-level and international tests) has come into prominence with worldwide testing such as TIMMS (Trends in International Mathematics and Science Study), PIRLS (Progress in International Reading Literacy) and PISA (Programme for International Student Assessment). The websites of these organizations carry a wealth of materials, information and access to test data (which many researchers use in their own work).

27.2 What are we testing?

Tests can concern achievement (what a person can do or knows), diagnosis (where the strengths and weaknesses of a student are; where the student is going wrong or having problems), aptitude (what the student is good at doing), proficiency, performance, speed, and so on. Tests can be used to compare students; to see if a student has achieved a particular fixed criterion (e.g. mastery tests, i.e. regardless of comparing with other students) (see below: norm-referencing and criterion-referencing); to see how quickly students can work (speed tests); to see what skills a student has mastered (e.g. a power test); to diagnose (e.g. difficulties and problems) etc.

Hambleton (2012, p. 242) identifies eight kinds of test:

- norm-referenced achievement and aptitude tests (commercially produced);
- criterion-referenced achievement tests (often commercially produced);
- classroom tests (produced by researchers, and intended for one-off use);
- performance tests;

- personality tests;
- attitude scales;
- interest inventories (often commercially produced);
- questionnaires.

However, this is just a starting point. There are myriad tests, to cover all aspects of a student's life and for all ages (young children to old adults), for example:

Ability	Learning disabilities
Achievement	Locus of control
Admission	Motivation and interest
Anxiety	Neuropsychological assessment
Aptitude	Performance
Attainment	Performance in school subjects
Attitudes and values	Personality
Behavioural disorders	Personality disorders
Competence-based assessment	Placement
Computer-based assessment	Potential
Creativity	Projective tests
Critical thinking	Psychomotor development
Cross-cultural adjustment	Reading readiness
Depression	Self-esteem
Diagnostic assessment	Sensory and perceptual tests
Diagnosis of difficulties	Social adjustment
Higher order thinking	Spatial awareness
Intelligence	Special abilities and disabilities
Interest inventories	Stress and burnout
Introversion and extraversion	University entrance
Language proficiency tests	Verbal and non-verbal reasoning

The *Handbook of Psychoeducational Assessment* (Saklofske *et al*., 2001) includes sections on ability assessment, achievement assessment, behaviour assessment, cross-cultural cognitive assessment and neuropsychological assessment. The *Handbook of Psychological and Educational Assessment of Children: Intelligence, Aptitude and Achievement* (Reynolds and Kamphaus, 2003) provides a clear overview of, inter alia:

- the history of psychological and educational assessment;
- a practical model of test development;
- legal and ethical issues in the assessment of children;
- measurement and design issues in the assessment of children;
- intelligence testing, both verbal and non-verbal;
- memory testing;
- neuropsychological and biological perspectives on the assessment of children;
- assessment of academic skills;
- criterion-referenced testing;
- diagnostic assessment;
- writing abilities and instructional needs;
- assessment of learning disabilities;
- bias in aptitude assessment;
- assessment of culturally and linguistically diverse children;
- assessment of creativity;
- assessment of language impairment;
- assessment of psychological and educational needs of children with severe mental retardation and brain injury;
- computer-based assessment.

As can be seen, there is a copious amount of assessment and testing material available and it covers a very wide spectrum of topics. *The Nineteenth Mental Measurements Yearbook* (Carlson and Geisinger, 2014) and *Tests in Print IX* (Anderson *et al.*, 2016) are useful sources of published tests, as are specific publishers and the website of BUROS (e.g. http://buros.org/test-reviews-information). In designing tests, Izard (2005) and Kline (2016) are useful sources.

There are several organizations that publish lists of tests and suppliers:

- The American Psychological Association: Finding Information about Psychological Tests (www.apa.org/science/programs/testing/find-tests.aspx)
- The British Psychological Society:
  - Psychological Testing (http://ptc.bps.org.uk/psychological-testing)
  - Psychological Testing Centre (http://ptc.bps.org.uk)
  - Psychological Test Collection (www.bps.org.uk/what-we-do/bps/history-psychology-centre/collections-and-archives/psychological-test-collection/psychological-test-collection)
  - Finding a Suitable Test (http://ptc.bps.org.uk/psychological-testing/psychological-testing/i-want-find-suitable-test-use)
- The Buros Center for Testing (www.unl.edu/buros)
- Healthy Place (US) (www.healthyplace.com/psychological-tests)
- Assessment Psychology online (www.assessmentpsychology.com/psychsites.htm).

Additionally there are countless websites that list psychological tests and we identify some of these on the companion website. Standard texts that detail copious tests, suppliers and websites include: Gronlund and

Linn (1990); Kline (2000, 2016); Loewenthal (2001); Saklofske *et al*. (2001); Reynolds and Kamphaus (2003); Gronlund and Brookhart (2008); Aiken (2003); Miller *et al*. (2012).

27.3 Parametric and non-parametric tests

Parametric tests are designed to represent the wide population, for example, of a country or age group. They make assumptions about the wider population and its characteristics, i.e. the parameters are known. They assume that:

- there is a normal curve of distribution of scores in the population (the bell-shaped symmetry of the Gaussian curve of distribution seen, for example, in standardized scores of IQ or the measurement of people's height or the distribution of achievement on reading tests in the population as a whole);
- the characteristics of the wider population are known, so that the parameters of each element of a test can be fairly sampled and controlled;
- there are continuous and equal intervals between the test scores, and, with tests that have a true zero (see Chapter 38), the opportunity for a score of, say, 80 per cent to be double that of 40 per cent; this differs from the ordinal scaling of rating scales discussed in connection with questionnaire design where equal intervals between each score cannot be assumed.

Parametric tests are usually published tests which are commercially available and which have been piloted, validated and standardized on a large and representative sample of the whole population. They usually arrive complete with backup data on sampling, reliability and validity statistics which have been computed in devising the tests. Working with these tests enables the researcher to use statistics applicable to interval and ratio levels of data.

On the other hand, non-parametric tests make few or no assumptions about the distribution of the population (the parameters of the scores) or the characteristics of that population. The tests do not assume a regular bell-shaped curve of distribution in the wider population; indeed the wider population is perhaps irrelevant as these tests are designed for a given specific population: a class in school, a chemistry group, a primary school year group. Because they make no assumptions about the wider population, the researcher must work with appropriate non-parametric statistics (see Chapter 38).

The attraction of non-parametric statistics is their utility for small samples because they do not make any assumptions about how normal, even and regular the distributions of scores will be. Non-parametric tests have the advantage of being tailored to particular institutional, departmental and individual circumstances. They offer teachers a valuable opportunity for quick, relevant and focused feedback on student performance.

Commercially produced parametric tests are more powerful than non-parametric tests because they not only derive from standardized scores but enable the researcher to compare sub-populations with a whole population (e.g. to compare the results of one school or district authority with the whole country, for instance in comparing students' performance in norm-referenced or criterion-referenced tests against a national average score in that same test). They enable the researcher to use high-level statistics in data processing (see Chapters 38–44) and to make *inferences* from the results. Because non-parametric tests make no assumptions about the wider population a different set of statistics is available to the researcher (see Part 5). These can be used in very specific situations – one class of students, one year group, one style of teaching, one curriculum area – and hence are valuable to teachers.

27.4 Diagnostic tests

Diagnostic tests are designed to identify particular strengths, weaknesses and problems in the aspect with which they are concerned (akin to going to a doctor with a medical complaint). Diagnostic tests identify needs, difficulties, successes and where problems arise. Whilst teachers are constantly diagnosing students' needs, difficulties, strengths, weaknesses and problems, there is a wide array of formal, standardized diagnostic tests available in the public domain, often with restricted access (e.g. to registered educational psychologists).

Diagnostic tests are often used as the foundation for formative planning, informing what action needs to be taken next (just as a doctor diagnoses an illness and then prescribes treatment). The two are different: diagnosis does not prescribe treatment, the educationist then has to decide what 'treatment' to administer.

27.5 Norm-referenced, criterion-referenced and domain-referenced tests

Norm-referenced tests

A norm-referenced test compares students' achievements relative to other students' achievements (e.g. a national test of mathematical performance or a test of intelligence which has been standardized ('normed') on a large and representative sample of students between

the ages of six and sixteen). For example, a commercially produced intelligence test has been standardized so that, for instance, a score of 100 is that of a notional 'average' student and a score of 120 describes a student who is notionally above average. The concept of 'average' only makes sense when it is derived from or used for a comparison of students.

Norm-referencing is based on a curve of distribution, which may be the bell-shaped, symmetrical Gaussian curve or a skewed curve (e.g. skewed in order to give more students higher grades, though this is highly contentious). Norm-referencing is designed to fit quotas for what proportion of students will be awarded grades/percentages (e.g. 5 per cent might be awarded a grade A; 20 per cent a grade B; 40 per cent a grade C; 20 per cent a grade D; 10 per cent a grade E; and 5 per cent a Fail). They enable students to be ranked: 'an order of merit' (Izard, 2005, p. 20).

In educational institutions, a norm-referenced assessment enables institutions to guarantee a proportion of high achievers (though it also guarantees a proportion of low achievers). It means that standards vary across classes/groups/years, as the comparators vary. For example, a grade A student in one year might be a grade B in another year, if that second year comprises high-scoring students, or a grade D student in one year might be a failing student in another year if that second year comprises higher-scoring students. It means that students might obtain a first-class degree one year when, in another, the same work would only be awarded a second-class degree. An 'A' in one year is no guide as to the meaning or standard of 'A' in another year; an 'F' in one year is no guide as to an 'F' in another. As Izard (2005) argues, the fair use of norm-referenced tests relies on the curriculum being static over time (p. 19).

However, the argument is also raised that students have a right to expect that the standard of their grading and awards will not be affected by the numbers of students in a course or programme or by quotas but by their performance on given, transparent criteria, and only against these criteria. Similarly, it is argued that students have a right to know at the start of their courses, i.e. in advance of the course, what constitutes an 'A' grade, what constitutes a 'B' grade and so on – this concerns equity and transparency, rather than being told that their grades will only be awarded once the number of students on the course and their marks have been calculated, or that there is a limit to the number/proportion of people at each grade.

Just as a norm-referenced system guarantees a certain proportion of high grades, for example, A and B, so, by definition, it also guarantees a proportion of low grades and failures, regardless of actual performance, in order to conform to the curve of distribution.

The educational defensibility or desirability of norm-referencing may be questionable: a 'good' student may end up failing or scoring poorly if the class or group of students with whom she/he is being compared is even better. Norm-referencing may be useful for selection (e.g. for a fixed, limited number of places in an elite university), but it may not be equitable.

Further, norm-referenced tests (e.g. standardized tests) used in worldwide testing assume that the curriculum of each country is the same, and this may not be the case, for example, algebra and geometry are introduced much later in an American student's school curriculum than in an Australian curriculum (Izard, 2005).

Criterion-referenced tests

A criterion-referenced test does not compare student with student but, rather, requires the student to fulfil a given set of criteria, a predefined and absolute standard or outcome (Cunningham, 1998), for example, in terms of knowledge or skills. It tests what a student can and cannot do, and what he/she knows and does not know, regardless of what any other students can and cannot do. For example, a driving test is usually criterion-referenced since to pass it requires the ability to meet certain test items – reversing round a corner, undertaking an emergency stop etc. – *regardless* of how many other people have or have not passed the driving test. Similarly, many tests of playing a musical instrument require specified performances, for example, the ability to play a particular scale or arpeggio, the ability to play a Bach fugue without hesitation or technical error. If the student meets the criteria, then he or she passes the examination.

In criterion-referenced assessment the specific criteria for success are set out in advance and students are assessed on the extent to which they have achieved them, without any reference being made to the achievements of other students (which is norm-referencing). If they meet the criteria then they achieve the grade, regardless of how many other students do or do not achieve the grade (i.e. unlike norm-referencing).

There are minimum competency cut-off levels, below which students are deemed not to have achieved the criteria, and above which different grades or levels can be awarded for the achievement of criteria – for example, a grade A, B, C etc.

In a criterion-referenced assessment, unlike in a norm-referenced assessment, there are no ceilings on the numbers of students who might be awarded a particular grade. In a norm-referenced system there might be only a small percentage who are able to achieve a

grade A because of the imposed proportion/quota (the 'norming' of the test), whereas in a criterion-referenced assessment, if everyone meets the criterion for a grade A then everyone is awarded a grade A, and if everyone should fail then everyone fails.

A criterion-referenced test provides the researcher with information about exactly what a student has learned, what she can do, whereas a norm-referenced test can only provide the researcher with information on how well one student has achieved in comparison to another, enabling rank orderings of performance and achievement to be constructed. Hence a major feature of the norm-referenced test is its ability to discriminate between students and their achievements – a well-constructed norm-referenced test enables differences in achievement to be measured acutely, i.e. to provide variability or a great range of scores. For a criterion-referenced test this is less of a problem; the intention here is to indicate whether students have achieved a set of given criteria, regardless of how many others might or might not have achieved them, hence variability or range is less important here. Criterion-referencing is often used in outcomes-based education.

Domain-referenced tests

An outgrowth of criterion-referenced testing is domain-referenced tests. Here importance is accorded to the careful and detailed specification of the content or the domain to be assessed. The domain is the particular field or area of the subject that is being tested, for example, light in science, two-part counterpoint in music, parts of speech in English. The domain is set out very clearly and very fully, such that the full depth and breadth of the content are established. Test items are then selected from this full field, with careful attention to sampling procedures so that representativeness of the wider field of items is ensured in the test items. The student's achievements on that test are computed to yield a proportion of the maximum score possible, and this, in turn, is used as an index of the proportion of the overall domain that she has grasped. So, for example, if a domain has 1,000 items and the test has 50 items from this 1,000, and the student scores 30 marks from the possible 50 then it is inferred that she has grasped 60 per cent ($\{30 \div 50\} \times 100$) of the domain of 1,000 items. Here inferences are made from a limited number of items to the student's achievements in the whole domain; this requires careful and representative sampling procedures for test items.

27.6 Commercially produced tests and researcher-produced tests

There is a battery of tests in the public domain which cover a vast range of topics and which can be used for evaluative purposes (some are indicated at the start of this chapter).

Most schools will have used published tests at one time or another. There are several attractions to using published tests:

- they are objective;
- they have been piloted and refined;
- they have been standardized across a named population (e.g. a region of the country, the whole country, a particular age group or various age groups) so that they represent a wide population;
- they declare how reliable and valid they are (mentioned in the statistical details which are usually contained in the manual of instructions for administering the test);
- they tend to be parametric tests, hence enabling sophisticated statistics to be calculated;
- they come complete with instructions for administration;
- they are often straightforward and quick to administer and mark;
- guides to the interpretation of the data are usually included in the manual;
- researchers are spared the task of having to devise, pilot and refine their own test.

On the other hand, commercially produced tests are expensive to purchase and administer; some are often targeted to special rather than to general populations (e.g. in psychological testing), and some may not be exactly suited to the purpose required (Howitt and Cramer, 2014). Further, several commercially produced tests have restricted release or availability, and the researcher might have to register with a particular association or be given clearance to use the test or to have copies of it. There are different levels of clearance, and certain parties or researchers may not be eligible to have a test released to them because they do not fulfil particular criteria for eligibility.

Published tests, by definition, are not tailored to institutional or local contexts or needs; indeed their claim to objectivity is made on the grounds that they are deliberately supra-institutional. The researcher wishing to use published tests must be certain that the purposes, objectives and content of the published tests match those of the research. For example, a published diagnostic test might not fit the needs of the research to

have an achievement test; a test of achievement might not have the predictive quality which the researcher seeks in an aptitude test; a published reading test might not address the areas of reading that the researcher is wishing to cover; a verbal reading test written in English might contain language which is difficult for a student whose first language is not English. These are important considerations. A text on evaluating the utility for researchers of commercially available tests is produced by the American Psychological Association (2014) in the *Standards for Educational and Psychological Testing* (www.apa.org/science/programs/testing/standards.aspx).

The golden rule for deciding whether to use a published test is that it must demonstrate *fitness for purpose*. If it fails to demonstrate this, then tests will have to be devised by the researcher. The attraction of this latter type is that a 'home-grown' test will be closely tailored to the local and institutional context, i.e. the purposes, objectives and content of the test will be deliberately fitted to the *specific* needs of the researcher in a specific, given context. Cronbach (1949), Gronlund and Linn (1990) and Miller *et al*. (2012) set out a range of criteria against which a commercially produced test can be evaluated for its fitness for purpose.

Researchers should be cautious in considering whether to employ commercially produced tests, particularly if using them with individuals and groups which are different from those in which the test was devised, as many tests show cultural bias (personality tests are prone to this, some being based on the 'Big-5' personality attributes which may exist but be less prominent in cultures other than those in which the test was produced).

Further, there is the issue of the language medium of the test; for example, using the Wechsler tests of intelligence (in English medium) with students who are not native speakers of English or who do not know about certain aspects of English culture renders the test less a test of intelligence and more a test of English-language ability and English cultural knowledge.

Many commercially produced tests are available in languages other than the original, but the translations should be checked to see if they are correct and for the cultural significance of the test items themselves, to see that they hold the same meaning, connotations and significance in the target language as they do in the original language. It is often dangerous to import tests developed in one language and one culture into another language and another culture, as there are problems of validity, bias, meaningfulness and reliability.

However, there are also several important considerations in devising a 'home-grown' test. Not only might it be time-consuming to devise, pilot, refine and then administer the test, but, because much of it will probably be non-parametric, there will be a more limited range of statistics which may be applied to the data than in the case of parametric tests.

The scope of tests and testing is far-reaching; no areas of educational activity are untouched by them. Achievement tests, largely summative in nature, measure achieved performance in a given content area. Aptitude tests are intended to predict capability, achievement potential, learning potential and future achievements. However, the assumption that these two constructs – achievement and aptitude – are separate is questionable (Cunningham, 1998); indeed often aptitude in, say, geography, at a particular age or stage will be measured by using an achievement test at that age or stage. Cunningham (1998) has suggested that an achievement test might include more straightforward measures of basic skills whereas aptitude tests might put these in combination, for example, combining reasoning (often abstract) and particular knowledge, i.e. achievement and aptitude tests differ according to what they are testing.

Not only do the tests differ according to what they measure, but, since both can be used predictively, they differ according to what they might be able to predict. For example, because an achievement test is often tied to a specific content area, it will be useful as a predictor of future performance in that content area but will be largely unable to predict future performance outside that content area. An aptitude test tends to test more generalized abilities (e.g. aspects of 'intelligence', skills and abilities that are common to several areas of knowledge or curricula), hence it can be used as a more generalized predictor of achievement. Achievement tests, Gronlund (1985) and Gronlund and Brookhart (2008) suggest, are more linked to school experiences, whereas aptitude tests encompass out-of-school learning and wider experiences and abilities. However, Cunningham (1998), arguing that there is a considerable overlap between the two types, suggests that the difference is largely cosmetic. An achievement test tends to be much more specific and linked to instructional programmes and cognate areas than an aptitude test, which looks for more general aptitudes (Hanna, 1993), for example, intelligence or intelligences (Gardner, 1993).

27.7 Constructing and validating a test

Researchers considering constructing a test of their own must be aware of classical test theory (CTT) and Item Response Theory (IRT). Classical test theory assumes

that there is a 'true score', which is the score which an individual would obtain on that test if the measurement was made without error and the expected score that would be gained over an infinite number of independent test administrations. It is the score that would be found by calculating the mean score that the individual test-taker would obtain on that same test if that person took it on an infinite number of occasions.

However, CTT recognizes that, in fact, errors do arise in the real world, due to, for example, cultural and socio-economic backgrounds and bias in the test, administration and marking of the test, and attitudes to the test by the test-takers. Hence tests provide an 'observed score' rather than a 'true score'; the observed score (X) is the true score (T) plus the error (E) (X=T+E). A true score in CTT depends on the contents of the test rather than the characteristics of the test-taker, and the difficulty of the items might depend on the characteristics of the sample (a sampling issue) rather than on the item itself, i.e. it may be difficult to compare the results of different test-takers on different tests. Readers are advised to review classical test theory and reliability in connection with this formula and the calculation of the error (e.g. Kline, 2005b).

By contrast, Item Response Theory (IRT) is based on the principle that it is possible to measure single, specific latent traits, abilities and attributes that, themselves, are not observable, i.e. to determine observable quantities of unobservable qualities (e.g. Hambleton, 1993). The theory/model assumes a relationship between a person's possession or level of a particular attribute, trait or ability and his/her response to a test item. IRT is also based on the view that it is possible:

- to identify objective levels of difficulty of an item, for example, the Rasch model (Wainer and Mislevy, 1990);
- to devise items that can discriminate effectively between individuals;
- to describe an item independently of any particular sample of people who might be responding to it, i.e. is not group dependent (the item difficulty and item discriminability are independent of the sample);
- to describe a testee's proficiency in terms of his or her achievement of an item of a known difficulty level;
- to describe a person independently of any sample of items that has been administered to that person (i.e. a testee's ability does not depend on the particular sample of test items);
- to specify and predict the properties of a test before it has been administered;
- for traits to be unidimensional (single traits are specifiable, e.g. verbal ability, mathematical proficiency) and to account for test outcomes and performance in terms of that unidimensional trait, i.e. for an item to measure a single, unidimensional trait;
- for a set of items to measure a common trait or ability;
- for a testee's response to any one test item not to affect his or her response to another test item;
- that the probability of the correct response to an item does not depend on the number of testees who might be at the same level of ability;
- to identify objective levels of difficulty of an item;
- to calculate a statistic which indicates the precision of the measured ability for each testee, and that this statistic depends on the ability of the testee and the number and properties of the test items.

In devising a test the researcher will have to consider not only the foundations of the test (e.g. in CTT or IRT) but also:

- the *purposes* of the test (for answering evaluation questions and ensuring that it tests what it is supposed to be testing, e.g. the achievement of the objectives of a piece of the curriculum);
- the *type* of test (e.g. diagnostic, achievement, aptitude, criterion-referenced, norm-referenced);
- the *objectives* of the test (cast in very specific terms so that the content of the test items can be seen to relate to specific objectives of a programme or curriculum);
- the *content* of the test (what is being tested and what the test items are);
- the relative *weightings* of the *content items* (to fit the *objectives* of the test, e.g. knowledge, understanding, application, synthesis, evaluation (using Bloom's *Taxonomy of Educational Objectives* (1956)));
- the relative *weightings* of the *content areas* (e.g. topic 1: 60 per cent; topic 2: 30 per cent; topic 3: 10 per cent);
- the relative *weightings* of the different *kinds of question* (e.g. multiple choice: 30 per cent; essay: 50 per cent; short answer: 20 per cent);
- the relative *weightings* of the items to match the *difficulty* of the test items (e.g. easy items: 10 per cent; slightly difficult: 20 per cent; moderately difficult: 40 per cent; difficult: 20 per cent; very difficult: 10 per cent);
- the *construction* of the test, involving *item analysis* in order to clarify the *item discriminability* and *item difficulty* of the test (see below);

- the *task suitability* of the test, for example, suitability for age and experiences of the students, what is being measured, how a question is to be answered;
- the *format* of the test: its layout, instructions, method of working and of completion (e.g. oral instructions to clarify what students must write, or a written set of instructions to introduce a practical piece of work);
- the *piloting* of the test;
- the *validity and reliability* of the test;
- the *scoring* of the test (allocation for marks, and on what criteria);
- the provision of a *manual of instructions* for the administration, marking and data treatment of the test (this is particularly important if the test is not to be administered by the researcher or if the test is to be administered by several different people, so that reliability is ensured by having a standard procedure).

Izard (2005, p. 33) suggests a sequence of test construction which proceeds thus: decisions to gather evidence and allocate resources → content analysis → item writing → by item review → item scoring → producing trial test and testing it → second item review followed by amendment → consideration of whether more items are required → production of final version of the test. In planning a test the researcher can proceed as set out below.

Identify the purposes of the test

The purposes of a test are several, for example to *diagnose* a student's strengths, weaknesses and difficulties, to measure *achievement*, to measure *aptitude* and *potential*, to assess *personality* attributes or types, to identify *readiness* for a programme (Gronlund and Linn (1990) term this 'placement testing', normally designed to discover whether students have the essential prerequisites to begin a programme, e.g. knowledge, skills, understandings).

These types of test occur at different stages. For example, the placement test is conducted prior to the commencement of a programme, and identifies the initial or 'entry' abilities in a student. If the placement test is designed to assign students to tracks, sets or teaching groups (i.e. to place them into administrative or teaching groups), then the entry test might be criterion-referenced or norm-referenced; if it is designed to measure detailed starting points, knowledge, abilities and skills then it might be more criterion-referenced as it requires a high level of detail. It has its equivalent in 'baseline assessment' and is an important feature if one is to measure the 'value-added' component of teaching and learning: one can only assess how much a set of educational experiences has added value to the student if one knows that student's starting point, starting abilities and achievements.

Formative testing is undertaken during a programme, and is designed to monitor students' progress during that programme, to measure achievement of sections of the programme and to diagnose strengths and weaknesses so that action can be targeted. It is typically criterion-referenced.

Diagnostic testing is an in-depth test to discover particular strengths, weaknesses and difficulties that a student is experiencing, and is designed to expose causes and specific areas of weakness or strength. This often requires the test to include several items about the same feature, so that, for example, several types of difficulty in a student's understanding are exposed; the diagnostic test requires test items that focus on each of a range of very specific difficulties that students might be experiencing, in order to identify the exact problems that they are having from a range of possible problems. Clearly this type of test is criterion-referenced.

Summative testing is the test given at the end of the programme, and is designed to measure achievement, outcomes or 'mastery'. This might be criterion-referenced or norm-referenced, depending to some extent on the use to which the results will be put (e.g. to award certificates or grades, to identify achievement of specific objectives, to control entry to university).

Identify the test specifications

Test specifications include:

- the programme objectives and student learning outcomes to be addressed;
- the content areas to be addressed;
- the relative weightings, balance and coverage of items, with weightings addressing objectives, content areas, kinds of question and difficulty of the items;
- the total number of items in the test;
- the number of questions required to address a particular element of a programme or learning outcomes;
- the exact items in the test.

To ensure validity in a test it is essential that the objectives of the test are addressed fairly in the test items. Objectives, it is argued (Mager, 1962; Wiles and Bondi, 1984, 2014), should: (a) be specific and expressed with an appropriate degree of precision; (b) represent intended learning outcomes; (c) identify the actual and observable behaviour which demonstrates achievement; (d) include an active verb; (e) be unitary (focusing on one item per objective).

The test must measure what it purports to measure. It should demonstrate several forms of validity (e.g. construct, content, concurrent, predictive, criterion-related), discussed below (see also Chapter 14).

One way of ensuring that the objectives are fairly addressed in test items is through a matrix frame that indicates the *coverage* of content areas, *objectives* of the programme and the *relative weighting* of the items on the test. Such a matrix is set out in Table 27.1, taking the example from a secondary school history syllabus.

Table 27.1 indicates the main areas of the programme to be covered in the test (*content areas*); then it indicates which *objectives* and detailed *content areas* are covered (1a – 3c) – these numbers refer to the identified specifications in the syllabus; then it indicates the marks/percentages to be awarded for each area. This indicates several points:

- the least emphasis is given to the build-up to and end of the war (10 marks each in the 'total' column);
- the greatest emphasis is given to the invasion of France (35 marks in the 'total' column);
- there is fairly even coverage of the objectives specified (the figures in the 'total marks possible' row only vary from 9 to 13);
- greatest coverage is given to objectives 2a and 3a, and least coverage is given to objective 1c;
- some content areas are not covered in the test items (the blanks in the matrix).

We have here a test scheme that indicates relative weightings, coverage of objectives and content, and the relation between these two latter elements. Gronlund and Linn (1990) and Miller *et al*. (2012) suggest that relative weightings should be addressed by first assigning percentages at the foot of each column, then assigning percentages at the end of each row, and then completing each cell of the matrix within these specifications. This ensures that appropriate sampling and coverage of the items are achieved. The example of the matrix refers to specific objectives as column headings; of course these could be replaced by factual knowledge, conceptual knowledge and principles, and skills for each of the column headings. Alternatively, they could be replaced with specific aspects of an activity, for example (Cohen *et al.*, 2010, p. 411): designing a crane, making the crane, testing the crane, evaluating the results, improving the design. Indeed these latter could become content (row) headings, as shown in Table 27.2. Here practical skills carry fewer marks than recording skills (the column totals), and

TABLE 27.1 A MATRIX OF TEST ITEMS

Content areas	*Objective/area of programme content*			*Objective/area of programme content*			*Objective/area of programme content*			*Total*
Aspects of the 1939–45 war	1a	1b	1c	2a	2b	2c	3a	3b	3c	
The build-up to the 1939–45 world war	1	2		2	1	1	1	1	1	10
The invasion of Poland	2	1	1	3	2	2	3	3	3	20
The invasion of France	3	4	5	4	4	3	4	4	4	35
The allied invasion	3	2	3	3	4	3	3	2	2	25
The end of the conflict	2	1		1	1	1	2	2		10
Total	11	10	9	13	12	10	13	12	10	100

TABLE 27.2 COMPILING ELEMENTS OF TEST ITEMS

Content area	*Identifying key concepts and principles*	*Practical skills*	*Evaluative skills*	*Recording results*	*Total*
Designing a crane	2	1	1	3	7
Making the crane	2	5	2	3	12
Testing the crane	3	3	1	4	11
Evaluating the results	3		5	4	12
Improving the design	2	2	3	1	8
Total	12	11	12	15	50

making and evaluating carry equal marks (the row totals).

This exercise also indicates the number of items to be included in the test; for instance in the example of the history test (Table 27.1), the matrix is $9 \times 6 = 54$ possible items, and in the 'crane' activity (Table 27.2) the matrix is $5 \times 4 = 20$ possible items. Of course, there could be considerable variation in this, for example more test items could be inserted if it were deemed desirable to test one cell of the matrix with more than one item (possible for cross-checking), or indeed there could be fewer items if a single test item could serve more than one cell of the matrix. The difficulty in matrix construction is that it can easily become a runaway activity, generating very many test items and, hence, leading to an unworkably long test; typically the greater the degree of specificity required, the greater the number of test items there will be. One skill in test construction is to be able to have a single test item that provides valid and reliable data for more than a single factor/area.

Having undertaken the test specifications, the researcher should have achieved clarity on: (a) the exact test items that test specified aspects of achievement of objectives, programmes, contents etc.; (b) the coverage and balance of coverage of the test items; and (c) the relative weightings of the test items.

Address validity and reliability

Validity

Validity concerns the extent to which the test tests what it is supposed to test; it must measure what it purports to measure. A test should demonstrate several kinds of validity (see also Chapter 14):

- *face validity*: the test must appear to assess what it was intended to test;
- *construct validity*: the extent to which the test measures a particular construct, trait, behaviour, evidenced through convergent validity and discriminant, divergent validity, and by correlating the test with other published tests with the same purposes and similar contents. The test must provide a fair operationalization of the construct – often abstract – in question, for example, intelligence, creativity, spatial awareness, problem solving. This is often the most challenging kind of validity to address, not least because opinion is divided on what a fair construction of the construct actually is. For example, exactly what intelligence is, and what proxy indicators of intelligence might be, can founder if there is disagreement on whether it is a single ability, a multiple ability (e.g. Gardner's 'multiple intelligences'), a composite, fixed or capable of being developed (nature or nurture). One statistical means of addressing construct validity is to seek inter-correlations between several items which are intended to measure the same construct (or to undertake factor analysis, itself based on inter-correlations). The principle here is that inter-correlations between items in a test that are intended to measure the same construct should be higher than inter-correlations between items that are not intended to measure the same construct or which are intended to measure different constructs. Different types of question intended to measure the same construct should have stronger inter-correlations than inter-correlations using the same types of question to assess different constructs;
- *content validity*: adequate and representative coverage of the domain, field, tasks, behaviours, knowledge etc., without interference from extraneous variables. The test must cover the intended contents in sufficient depth and breadth so as to be fair and adequate, and not to exceed the boundaries of content (i.e. not to cover items or contents that were not included in the programme or curriculum);
- *concurrent validity*: the extent to which the test scores correlate with those of other tests in a similar field;
- *predictive validity*: that the test accurately predicts final scores/outcomes. This concerns how much the results of an assessment can be used to predict achievements in the future, for example, how much the scores at university entrance level might be fair indicators of future degree classifications. Low predictive validity (e.g. lower than 50 per cent) suggests that limited credence should be placed in such uses;
- *criterion-related validity*: the extent to which the performance on the test enables the researcher to infer the individual's performance on a particular criterion of interest. This is often calculated as a correlation between the score on a test and a score in another indication of the item that the test was intended to measure, for example, a test of performance on a job-specific matter and the individual's actual performance on that job-specific item in the real situation;
- *cultural validity*: fairness to the language and culture of the individual test-takers, and avoidance of cultural bias: a feature of all research instruments, not solely tests;
- *consequential validity*: the results of the test are used fairly and ethically (discussed later), and are

only used for the purpose of, and ways in which, the test was constructed and intended to be used. This requires the researcher to be clear on the intentions of the test and its uses.

Reliability

Reliability concerns the degree of confidence that can be placed in the results, which is often a matter of statistical calculation and subsequent test redesign. It concerns the stability and consistency of test scores (e.g. if a student takes a test twice, or similar versions of the same test, the scores should be similar). Reliability is addressed through the forms and techniques set out in Chapter 14: test–re-test, parallel forms, split-half and internal consistency (Cronbach's alpha).

Reliability is compromised when students of the same ability and achievements score different results on the same test, when the same student scores differently on different tests of the same matters/contents, or when the same student scores differently on the same (or very similar) test on a different occasion. Reliability means that the results are consistent and reproducible with different markers, occasions, test items, test types, marking conventions, grading procedures and contexts.

Reliability concerns consistency and dependability, for example, of marking practices/conventions and standards. An assessment has little reliability if it yields different results in the hands of another assessor or different results for similar students. Reliability requires comparability of practices to be addressed. This can be undertaken prior to assessments by agreement trials, so that a range of assessors can be clear on, and can agree on, the specific marks and grades to be awarded for particular samples of work, examination scripts, course work and marks scored in each element of an overall assessment, though in practice it often only becomes an issue in the post-assessment agreement of marks (moderation) and awards. Reliability concerns, and affects, the degree of confidence that one can put on assessment data and their interpretation.

Not only must reliability be addressed but it must be *seen* to be addressed; marking must be seen to be fair, i.e. transparency. Reliability can be improved by, among other things: joint planning between researchers/markers; using the intended learning outcomes to agree objectives for the test; and developing common activities focused on agreed objectives.

Sources of unreliability

There are many threats to reliability, and researchers (and teachers) need to do their best to reduce them.

With respect to *examiners* and *markers*:

- errors in marking (e.g. attributing, adding and transfer of marks);
- inter-rater reliability (different markers giving different marks for the same or similar pieces of work);
- inconsistency in the examiner/marker (e.g. being harsh in the early stages of the marking and lenient in the later stages of the marking of many scripts);
- variations in the award of grades for work that is close to grade boundaries (some markers placing the score in a higher or lower category than other markers);
- the halo effect (or its opposite, the horns effect), wherein a student who is judged to do well or badly in one test/assessment is given undeserved favourable or unfavourable test/assessment respectively in other areas.

With reference to the *students* and *teachers* themselves:

- motivation and interest in the task has a considerable effect on performance. Motivation to participate in tests is strongest when students have been helped to see its purpose, and where the examiner maintains a warm, purposeful attitude towards them during the testing session;
- the relationship (positive to negative) between the tester and the testee exerts an influence on the test. This takes on increasing significance where the students know the researcher/teacher personally and professionally – and vice versa – and where the assessment situation involves face-to-face contact between the researcher/teacher and the student;
- the conditions – physical, emotional, social – exert an influence on the test, particularly if they are unfamiliar. The advice generally given in connection with the location of a test is that the test-room should be well-lit, quiet and adequately ventilated. Wherever possible, students should take tests in familiar settings, preferably in their own classrooms under normal school conditions. Distractions in the form of extraneous noise, walking about the room by the invigilator and intrusions into the room can have an impact on the scores of the test-takers;
- the Hawthorne effect, wherein, in this context, simply informing a student that this is a test will be enough to disturb her performance – for better or worse (either case not being a fair reflection of her usual abilities);
- teacher's marking may be prone to bias in the halo effect and in the teacher's own confusion – perhaps with the best of motives – between effort and

achievement, rewarding high effort and industry even though the achievement may be poor (and blind marking may not be as 'blind' as is imagined, as teachers recognize their students' handwriting);
- distractions (including superfluous information);
- the time of the day, week, month can exert an influence on performance. Some students are fresher in the morning and more capable of concentration;
- students are not always clear on what they think is being asked in the question; they may know the right answer but not infer that this is what is required in the question;
- a student may perform better with a different set of questions which test the same matters;
- teachers teach to the test. This is perhaps unsurprising in high-stakes tests. Here a biased result is obtained: students do well in the test without much understanding; they are groomed in test-taking;
- teachers and students practise test-like materials;
- teachers conducting their own tests may resort unconsciously to simplistic testing rather than richer and more extended forms of assessment;
- a student may be able to perform a specific skill in a test but not be able to select or perform it in the wider context of learning;
- cultural, ethnic and gender background affect how meaningful a test task or activity is to students, and meaningfulness affects their performance;
- students' personalities may make a difference to their test performance;
- students' learning strategies and styles may make a difference to their test performance;
- marking practices are not always reliable; teachers may be too generous, marking by effort and ability rather than performance;
- the context in which the task is presented affects performance: some students can perform the task in everyday life but not under test conditions.

With regard to the *test items*:

- the task itself may be multi-dimensional, for example, testing 'reading' may require several components and constructs. Students can execute a mathematics operation in the mathematics class but they cannot perform the same operation in, for example, a physics class. This raises the issue of the number of contexts in which the behaviour must be demonstrated before a criterion is deemed to have been achieved. The *context* of the task affects the student's performance;
- the validity of the items may be in question;
- the language of the test and the tester exerts an influence on the testee, for example, if the test is conducted in the testee's second language;
- the readability level of the task can exert an influence on the test, for example, a difficulty in reading might distract from the purpose of a test which is to test the use of a mathematical algorithm;
- the number and type of operations and stages to a task: a student might know how to perform each element, but when they are presented in combination the size of the task can be overwhelming;
- the form and presentation of questions affects the results, giving variability in students' performances;
- a single error early on in a complex sequence may confound the later stages of the sequence (within a question or across a set of questions), even though the student might have been able to perform the later stages of the sequence, thereby preventing the student from gaining credit for all she or he can, in fact, do;
- essay questions may favour males if they concern impersonal topics and females if they concern personal and interpersonal topics;
- males may perform better than females on multiple-choice questions and females may perform better than males on essay-type questions, and females may perform better in written work than males;
- some students may be more anxious about tests than others, and consequently their performance may suffer;
- questions and tests may be culture-bound: what is comprehensible in one culture may be incomprehensible in another;
- the test may be so long, in order to ensure coverage, that boredom and loss of concentration may impair reliability.

Select the contents of the test

Item analysis

Gronlund and Linn (1990), Izard (2005) and Miller *et al*. (2012) suggest that an item analysis will need to consider:

- the suitability of the format of each item for the (learning) objective (appropriateness);
- the fairness of the item for the age, educational level and experiences of the student;
- the representativeness of the item (an individual item or group of items) for the matter(s) to be tested;
- the ability of each item to enable students to demonstrate their performance of the (learning) objective (relevance);

- the task requirements and contents;
- what the task is intended to cover (e.g. which elements of the curriculum);
- the clarity of the task and its requirements for each item;
- the nature and contents of the answer (e.g. a single correct answer, a 'best' answer for the item);
- the language and wording used (e.g. its difficulty and age-appropriateness);
- the straightforwardness of the task;
- the removal of unintended clues;
- the unambiguity of the outcome of each item, and agreement on what that outcome should be;
- the cultural, gender, racial etc. fairness of each item;
- the meaningfulness of the task (e.g. for age, ability, experiences, culture);
- the independence of each item (i.e. where the influence of other items of the test is minimal and where successful completion of one item is not dependent on successful completion of another);
- the adequacy of coverage of each (learning) objective by the items of the test;
- practical concerns, such as timing and duration; whether the students will be told the scoring of items; advice on how and where to answer, for example, on the examination question sheet, on separate paper/computer; layout of the paper.

In test construction the researcher will need to consider how each element to be tested will be *operationalized*: (a) what indicators and kinds of evidence of achievement of the objective will be required; (b) what indicators of high, moderate and low achievement there will be; (c) how the task will be introduced (e.g. written, oral, pictorial, computer, practical demonstration); (d) what the students will be doing when they are working on each element of the test; (e) what the outcome of the test will be (e.g. a written response, a tick in a box of multiple-choice items, an essay, a diagram, a computation).

The Task Group on Assessment and Testing in the UK (1988) suggest that attention will have to be given to the *presentation*, *operation* and *response* modes of a test: (a) how the task will be introduced (e.g. oral, written, pictorial, computer, practical demonstration); (b) what the students will be doing when they are working on the test (e.g. mental computation, practical work, oral work, written); and (c) what the outcome will be – how they will show achievement and present the outcomes (e.g. choosing one item from a multiple-choice question, writing a short response, open-ended writing, oral, practical outcome, computer output).

Operationalizing a test from objectives can proceed by stages:

- identify the objectives/outcomes/elements to be covered;
- break down the objectives/outcomes/elements into constituent components or elements;
- select the components that will feature in the test, such that, if possible, they will represent the larger field (i.e. domain-referencing, if required);
- recast the components in terms of specific, practical, observable behaviours, activities and practices that fairly represent and cover that component;
- specify the kinds of data required to provide information on the achievement of the criteria;
- specify the success criteria (performance indicators) in practical terms, working out marks and grades to be awarded and how to address weightings;
- write each item of the test;
- conduct a pilot to refine the language/readability and presentation of the items, to gauge item discriminability, item difficulty and distracters (discussed below), and to address validity and reliability.

Item analysis, Gronlund and Linn (1990, p. 255) aver, is designed to ensure that: (a) the items function as they are intended, for example, that criterion-referenced items fairly cover the fields and criteria and that norm-referenced items demonstrate *item discriminability* (discussed below); (b) the level of difficulty of the items is appropriate (see below: *item difficulty*); (c) the test is reliable (free of distractors – unnecessary information and irrelevant cues, see below: *distractors*) (see Millman and Greene, 1993). An item analysis will consider the accuracy levels available in the answer, the item difficulty, the importance of the knowledge or skill being tested, the match of the item to the programme and the number of items to be included. The foundation for item analysis lies in Item Response Theory, discussed earlier.

Item discriminability

In constructing a test the researcher will need to address the *item discriminability* of each item of the test. *Item discriminability* refers to the potential of the item in question to be answered correctly by those students who have a lot of the particular quality that the item is designed to measure and to be answered incorrectly by those students who have less of the particular quality that the same item is designed to measure. In other words, how effective is the test item in showing up differences among a group of students? Does the item enable us to discriminate between students' abilities in

a given field? An item with high discriminability will enable the researcher to see a potentially wide variety of scores on that item; an item with low discriminability will show scores on that item poorly differentiated. Clearly a high measure of discriminability is desirable, and items with low discriminability should be discarded.

Suppose the researcher wishes to construct a test of mathematics for eventual use with thirty students in a particular school (or with class A in a particular school). The researcher devises a test and *pilots* it in a different school or class B respectively, administering the test to thirty students of the same age (i.e. she matches the sample of the pilot school or class to the sample in the school which eventually will be used). The scores of the thirty pilot children are then split into three groups of ten students each (high, medium and low scores). It would be reasonable to assume that there will be more correct answers to a particular item among the high scorers than among the low scorers. For each item compute the following:

$$\frac{A-B}{\frac{1}{2}(N)}$$

where:
A = the number of *correct* scores from the high-scoring group;
B = the number of *correct* scores from the low-scoring group;
N = the *total* number of students in the two groups.

Suppose all ten students from the high-scoring group answered the item correctly and two students from the low-scoring group answered the item correctly. The formula would work out thus:

$$\frac{8}{\frac{1}{2}(10+10)} = 0.80 \text{ (index of discriminability)}$$

The maximum index of discriminability is 1.00. Any item whose index of discriminability is lower than 0.67 is too undiscriminating and should be reviewed to find out whether this is due to ambiguity in the wording or possible clues in the wording. If this is not the case, then whether the researcher uses an item with an index lower than 0.67 is a matter of judgement. The item in the example here would be appropriate to use in a test. For a further discussion of item discriminability, see Linn (1993) and Aiken (2003).

One can use the discriminability index to examine the effectiveness of *distractors*. This is based on the premise that an effective distractor should attract more students from a low-scoring group than from a high-scoring group. Consider the following example, where low- and high-scoring groups are identified:

	A	*B*	*C*
Top 10 students	10	0	2
Bottom 10 students	8	0	10

In example A, the item attracts only a few more correct responses (10) from the top ten students than the bottom ten (8) and hence is a poor distractor. Example B is an ineffective distractor because nobody was included from either group. Example C is an effective distractor because it includes far more students from the bottom ten students (10) than the higher group (2). However, in this case any ambiguities must be ruled out before the discriminating power can be improved.

Distractors

Distractors are the stuff of multiple-choice items, where incorrect alternatives are offered, and students have to select the correct alternatives. Here a simple frequency count of the number of times a particular alternative is selected will provide information on the effectiveness of the distractor: if it is selected many times then it is working effectively; if it is seldom or never selected then it is not working effectively and should be replaced.

Item difficulty

Researchers do not wish to have a test which is too easy (the ceiling effect) nor too difficult (the floor effect) (Ary *et al*., 2002, pp. 218–19). In constructing a test, the researcher will need to address the *item difficulty* of each item of the test. If we wish to calculate the *item difficulty* of a test, we can use the following formula:

$$\frac{A}{N} \times 100$$

where:
A = the number of students who answered the item correctly;
N = the *total* number of students who attempted the item.

Hence if twelve students out of a class of twenty answered the item correctly, then the formula would work out thus:

$$\frac{12}{20} \times 100 = 60\%$$

The maximum index of difficulty is 100 per cent. Items falling below 33 per cent and above 67 per cent are likely to be too difficult and too easy respectively. It would appear, then, that this item would be appropriate to use in a test. Here, again, whether the researcher uses an item with an index of difficulty below or above the cut-off points is a matter of judgement. In a norm-referenced test the item difficulty should be around 50 per cent (Frisbie, 1981). For further discussion of item difficulty, see Linn (1993) and Hanna (1993).

With regard to item difficulty, in a criterion-referenced test the level of difficulty is that which is appropriate to the task or objective. Hence if an objective is easily achieved then the test item should be easily achieved; if the objective is difficult then the test item should be correspondingly difficult. This means that, unlike a norm-referenced test where an item might be reworked in order to increase its discriminability index, this is less of an issue in criterion-referencing. Of course, this is not to deny the value of undertaking an item difficulty analysis; rather it is to question the centrality of such a concern. Gronlund and Linn (1990, p. 265) suggest that where instruction has been effective the item difficulty index of a criterion-referenced test will be high.

Given that the researcher can only know the degree of item discriminability and difficulty once a test has been undertaken, there is an unavoidable need to pilot home-grown tests. Items with limited discriminability and limited difficulty must be weeded out and replaced, those items with the greatest discriminability and the most appropriate degrees of difficulty can be retained; this can only be undertaken once data from a pilot have been analysed.

Item discriminability and item difficulty have different significance in norm-referenced and criterion-referenced tests. In a norm-referenced test we wish to compare students with each other, hence item discriminability is very important. In a criterion-referenced test, on the other hand, it is not important per se to be able to compare or discriminate between students' performance. For example, it may be the case that we wish to discover whether a group of students has learnt a particular body of knowledge, that is the objective, rather than, say, finding out how many have learned it better than others. Hence it may be that a criterion-referenced test has very low discriminability if all the students achieve very well or achieve very poorly, but the discriminability is less important than the fact that the students have or have not learnt the material. A norm-referenced test would regard such a poorly discriminating item as unsuitable for inclusion, whereas a criterion-referenced test might regard such an item as providing useful information (on success or failure).

In addressing item discriminability, item difficulty and distractor effects of particular test items, it is advisable to pilot these tests and to avoid placing too great a store on indices of difficulty and discriminability that are computed from small samples.

In constructing a test with item analysis, item discriminability, item difficulty and distractor effects in mind, it is important also to consider the actual requirements of the test (Nuttall, 1987; Cresswell and Houston, 1991), for example:

- Are all the items in the test equally difficult?
- Which items are easy, moderately hard, hard, very hard?
- What kinds of task is each item addressing, for example, is it: (a) repeating known knowledge; (b) applying known knowledge; (c) a synthesis item – bringing together and integrating diverse areas of knowledge?
- What makes some items more difficult than the rest?
- Are the items sufficiently within the experience of the students?
- How motivated will students be by the contents of each item (i.e. how relevant they perceive the item to be, how interesting it is)?

The contents of the test will also need to take account of the notion of *fitness for purpose*, for example in the types of test items. Here the researcher will need to consider whether ability, understanding and achievement will be best demonstrated in, for example (Lewis, 1974; Cohen *et al*., 2010, chapter 15):

- an open essay;
- a factual and heavily directed essay;
- short answer questions;
- divergent thinking items;
- completion items;
- multiple-choice items (with one correct answer or more than one correct answer);
- matching pairs of items or statements;
- inserting missing words/numbers;
- incomplete sentences or incomplete, unlabelled diagrams;
- true/false statements;
- short essay questions;
- long essay questions;
- open-ended questions where students are given guidance on how much to write (e.g. 300 words, a sentence, a paragraph);
- closed questions.

These items can test recall, knowledge, comprehension, application, analysis, synthesis and evaluation, i.e. different orders of thinking, and the weighting of marks will reflect the emphasis given to different levels (orders) of thinking: low-order to high-order thinking. These take their rationale from Bloom (1956) on hierarchies of thinking – from low-order (comprehension, application), through middle-order (analysis, synthesis) to higher-order thinking (evaluation, judgement, criticism, creation). The selection of the form of the test item will be based on the principle of gaining the maximum amount of information in the most economical way (and machine-scorable multiple-choice completion tests, for example, enable optical mark readers and scanners to enter and process large-scale data rapidly).

In considering the contents of a test, the test writer must also consider the *scale* for some kinds of test. Many psychological tests used in educational research are *unidimensional*, that is, the items all measure a single element or dimension. Other tests may be *multi-dimensional*, i.e. where two or more factors or dimensions are being measured in the same test. Test constructors must be clear whether they are using a unidimensional or a multi-dimensional scale. Many texts, whilst advocating the purity of using a unidimensional test that measures a single construct or concept, also recognize the efficacy, practicality and efficiency in using multi-dimensional tests. For example, though one might regard intelligence as a unidimensional factor, in fact a stronger measure of intelligence would be obtained by regarding it as a multi-dimensional construct, thereby requiring multi-dimensional scaling. Some items on a test are automatically unidimensional, for example, age, hours spent on homework.

Further, the selection of the items needs to be considered in order to have the highest reliability. Let us say that we have ten items that measure students' negative examination stress. Each item is intended to measure stress, for example:

Item 1 Loss of sleep at examination time;
Item 2 Anxiety at examination time;
Item 3 Irritability at examination time;
Item 4 Depression at examination time;
Item 5 Tearfulness at examination time;
Item 6 Unwillingness to do household chores at examination time;
Item 7 Mood swings at examination time;
Item 8 Increased consumption of coffee at examination time;
Item 9 Positive attitude and cheerfulness at examination time;
Item 10 Eager anticipation of the examination.

You run a reliability test (see Chapter 40) of internal consistency and find strong inter-correlations between items 1–5 (e.g. around 0.85), negative correlations between items 9 and 10 and all the other items (e.g. –0.79), and a very low inter-correlation between items 6 and 8 and all the others (e.g. 0.26). Item-to-total correlations (one kind of item analysis in which the item in question is correlated with the sum of the other items) vary here. What do you do? You can retain items 1–5. For items 9 and 10 you can reverse the scoring (as these items looked at positive rather than negative aspects), and for items 6 and 8 you can consider excluding them from the test, as they appear to be measuring something else. Such item analysis is designed to include items that measure the same construct and to exclude items that do not. We refer readers to Howitt and Cramer (2005, chapter 12) for further discussion of this.

An alternative approach to deciding which items to retain or exclude from the list of ten items above is to use factor analysis (see Chapter 43). Factor analysis groups together a cluster of similar items and keeps that cluster separate from clusters of other items. So, for our example above, the factor analysis could have found, by way of illustration, three factors:

- positive feelings (items 9 and 10);
- negative psychological states (items 2, 3, 4, 5, 7);
- physical, behavioural changes (items 1, 6, 8).

By looking at the factor loadings (see Chapter 43) the researcher would have to decide which were the most appropriate factors to retain, and thereby which items to include and exclude. As a general rule, items with low factor loadings (e.g. <0.3) should be considered for exclusion, as they do not contribute sufficiently to the factor. Factor analysis will indicate, also, whether the construct is unidimensional or multi-dimensional (if there is only one factor it is probably unidimensional).

Consider the form of the test

Much of the discussion in this chapter assumes that the test is of the pen-and-paper variety. Clearly this need not be the case: for example, tests can be written, oral, practical, interactive, computer-based, dramatic, diagrammatic, pictorial, photographic, involve the use of audio and video material, presentations, role-play and simulations. Oral tests, for example, can be conducted if the researcher feels that reading and writing will obstruct the true purpose of the test (i.e. it becomes a reading and writing test rather than, say, a test of mathematics). The form of the test will still need to consider, for example, reliability and validity, difficulty, discriminability, marking and grading, item analysis,

timing. Indeed several of these factors take on an added significance in non-written forms of testing; for example: (a) reliability is a major issue in judging live musical performance or the performance of a gymnastics routine – where a live, 'one-off' event takes place; (b) reliability and validity are significant issues in group performance or group exercises – where group dynamics may prevent a testee's true abilities from being demonstrated. Clearly the researcher will need to consider whether the test will be undertaken individually, or in a group, and what form it will take.

Write the test item

Here the test item is written which will test the knowledge, skills, aptitudes, performance etc. of the student. Care must be taken to ensure that only the area in question is included, not other areas, and that the item is unambiguous, indicating clearly what is required and what will provide evidence of the matter being tested. Care must be taken to ensure that what is required to be measured is included rather than, for example, only what is easily measured (Izard, 2005, p. 35), i.e. avoid superficiality.

The test will need to address the intended and unintended clues that might be provided in it, for example (Morris *et al*., 1987):

- the number of blanks might indicate the number of words required;
- the number of dots might indicate the number of letters required;
- the length of blanks might indicate the length of response required;
- the space left for completion will give cues about how much to write;
- blanks in different parts of a sentence will be assisted by the reader having read the other parts of the sentence (anaphoric and cataphoric reading cues).

Hanna (1993, pp. 139–41) and Cunningham (1998) provide several guidelines for constructing short-answer items to overcome some of these problems:

- make the blanks close to the end of the sentence;
- keep the blanks the same length;
- ensure that there can be only a single correct answer;
- avoid putting several blanks close to each other (in a sentence or paragraph) such that the overall meaning is obscured;
- only make blanks of key words or concepts, rather than of trivial words;
- avoid addressing only trivial matters;
- ensure that students know exactly the kind and specificity of the answer required;
- specify the units in which a numerical answer is to be given;
- use short answers for testing knowledge recall.

With regard to multiple-choice items there are several potential problems:

- the number of choices in a single multiple-choice item, and whether there is one or more right answer(s);
- the number and realism of the distractors in a multiple-choice item, for example, there might be many distractors but many of them are too obvious to be chosen – there may be several redundant items;
- the sequence of items and their effects on each other;
- the location of the correct response(s) in a multiple-choice item.

Gronlund and Linn (1990), Hanna (1993, pp. 161–75), Cunningham (1998) and Aiken (2003) set out several suggestions for constructing effective multiple-choice test items:

- ensure that they catch significant knowledge and learning rather than low-level recall of facts;
- frame the nature of the issue in the stem of the item, ensuring that the stem is meaningful in itself (e.g. replace the general stem 'sheep: (a) are graminivorous, (b) are cloven-footed, (c) usually give birth to one or two lambs at a time', with 'how many lambs are normally born to a sheep at one time?');
- ensure that the stem includes as much of the item as possible, with no irrelevancies;
- avoid negative stems to the item;
- keep the readability levels low;
- ensure clarity and unambiguity;
- ensure that all the options are plausible so that guessing of the only possible option is avoided;
- avoid the possibility of students making the correct choice through incorrect reasoning;
- include some novelty to the item if it is being used to measure understanding;
- ensure that there can only be a single correct option (if a single answer is required) and that it is unambiguously the right response;
- avoid syntactical and grammatical clues by making all options syntactically and grammatically parallel and by avoiding matching the phrasing of a stem with similar phrasing in the response;

- avoid including in the stem clues as to which may be the correct response;
- ensure that the length of each response item is the same (e.g. to avoid one long correct answer from standing out);
- keep each option separate, avoiding options which are included in each other;
- ensure that the correct option is positioned differently for each item (e.g. so that it is not always option 2);
- avoid using options such as 'all of the above' or 'none of the above' unless essential;
- avoid answers from one item being used to cue answers to another item – keep items separate.

The response categories of tests need to be considered, and we refer readers to our discussion of this topic in Chapter 24 on questionnaires (e.g. Likert scales, Guttman scales, semantic differential scales, Thurstone scales).

Morris *et al.* (1987, p. 161), Gronlund and Linn (1990), Hanna (1993, p. 147), Cunningham (1998) and Aiken (2003) also indicate particular problems in true/false questions:

- ambiguity of meaning;
- some items might be partly true or partly false;
- items that polarize: too easy or too hard;
- most items might be true or false under certain conditions;
- it may not be clear to the student whether facts or opinions are being sought;
- as this is dichotomous, students have an even chance of guessing the correct answer;
- an imbalance of true to false statements;
- some items might contain 'absolutes' which give powerful clues, for example, 'always', 'never', 'all', 'none'.

To overcome these problems the authors suggest several points that can be addressed:

- avoid generalized statements (as they are usually false);
- avoid trivial questions;
- avoid negatives and double negatives in statements;
- avoid over-long and over-complex statements;
- ensure that items are rooted in facts;
- ensure that statements can only be either true or false;
- write statements in everyday language;
- decide where it is appropriate to use 'degrees' – 'generally', 'usually', 'often' – as these are capable of interpretation;
- avoid ambiguities;
- ensure that each statement only contains one idea;
- if an opinion is to be sought then ensure that it is attributable to a named source;
- ensure that true statements and false statements are equal in length and number.

Morris *et al.* (1987), Hanna (1993, pp. 150–2), Cunningham (1998) and Aiken (2003) also indicate particular potential difficulties in matching items:

- it might be very clear to a student which items in a list simply *cannot* be matched to items in the other list (e.g. by dint of content, grammar, concepts), thereby enabling the student to complete the matching by elimination rather than understanding;
- one item in one list might be able to be matched to several items in the other;
- the lists might contain unequal numbers of items, thereby introducing distractors – rendering the selection as much a multiple-choice item as a matching exercise.

The authors suggest that difficulties in matching items can be addressed thus:

- ensure that the items for matching are homogeneous – similar – over the whole test (to render guessing more difficult);
- avoid constructing matching items to answers that can be worked out by elimination, for example, by ensuring that: (a) there are different numbers of items in each column so that there are more options to be matched than there are items; (b) students can avoid being able to reduce the field of options as they increase the number of items that they have matched; (c) the same option may be used more than once;
- decide whether to mix the two columns of matched items, i.e. ensure, if desired, that each column includes both items and options;
- sequence the options for matching so that they are logical and easy to follow, for example, by number, by chronology;
- avoid over-long columns and keep the columns on a single page;
- make the statements in the options columns as brief as possible;
- avoid ambiguity by ensuring that there is a clearly suitable option that stands out from its rivals;
- make it clear what the nature of the relationship should be between the item and the option (on what terms they relate to each other);
- number the items and letter the options.

With regard to essay questions, there are several claimed advantages. For example, an essay, as an open form of testing, enables complex learning outcomes to be measured, it enables the student to integrate, apply and synthesize knowledge, to demonstrate the ability for expression and self-expression, and to demonstrate higher-order and divergent cognitive processes. Further, it is comparatively easy to construct an essay title. On the other hand, essays have been criticized for yielding unreliable data (Gronlund and Linn, 1990; Cunningham, 1998) and for being prone to unreliable scoring (inconsistent and variable), neglectful of intended learning outcomes and prone to marker bias and preference (being too intuitive, subjective, holistic and time-consuming to mark). To overcome these difficulties, the authors suggest that:

- the essay question must be restricted to those learning outcomes that are unable to be measured more objectively;
- the essay question must ensure that it is clearly linked to desired learning outcomes; that it is clear what behaviours the students must demonstrate;
- the essay question must indicate the field and tasks very clearly (e.g. 'compare', 'justify', 'critique', 'summarize', 'classify', 'analyse', 'clarify', 'examine', 'apply', 'evaluate', 'synthesize', 'contrast', 'explain', 'illustrate');
- time limits are set for each essay;
- options are avoided, or, if options are given, ensure that, if students have a list of titles from which to choose, each title is equally difficult and equally capable of enabling the student to demonstrate achievement, understanding etc.;
- marking criteria are prepared and are explicit, indicating what must be included in the answers and the points to be awarded for such inclusions or ratings to be scored for the extent to which certain criteria have been met;
- decisions are agreed on how to address and score irrelevancies, inaccuracies, poor grammar and spelling;
- the work is blind double marked (markers are undisclosed to each other), and, where appropriate, blind marked (without the marker knowing (the name of) the essay writer).

These are issues of reliability (see Chapter 14). For a general introduction to writing test items, see Cohen and Wollack (2010).

Consider the layout of the test

This will include (Gronlund and Linn, 1990; Hanna, 1993; Linn, 1993; Cunningham, 1998):

- the nature, length and clarity of the instructions, for example, what to do, how long to take, how much to do, how many items to attempt, what kind of response is required (a single word, a sentence, a paragraph, a formula, a number, a statement etc.), how and where to enter the response, where to show the 'working out' of a problem, where to start new answers (e.g. in a separate booklet), whether one answer only is required to a multiple-choice item, or whether more than one answer is required;
- spreading out the instructions through the test, avoiding overloading students with too much information at first, and providing instructions for each section as they come to it;
- indicating the marks able to be awarded for each part of the test;
- minimizing ambiguity and taking care over the readability of the items;
- progression from the easy to the more difficult items of the test (i.e. the location and sequence of items);
- the visual layout of the page, for example, avoiding overloading students with visual material or words;
- the grouping of items – keeping together items that have the same contents or the same format;
- the layout of answer sheets/locations so that they can be entered onto computers and read by optical mark readers and scanners (if appropriate).

Layout can exert a profound effect on the test. The layout of the text should be such that it supports the completion of the test and that this is done as efficiently and as effectively as possible for the student.

Consider the timing of the test

This refers to two areas: (a) when the test will take place (the day of the week, month, time of day), and (b) the time allowances to be given to the test and its component items. With regard to the former, in part this is a matter of reliability, for the time of day, day of the week etc. might influence how alert, motivated or capable a student might be. With regard to the latter, the researcher will need to decide what time restrictions are being imposed and why, for example, is the pressure of a time constraint desirable – to show what a student can do under time pressure (a speed test) – or an unnecessary impediment, putting a time boundary around something that need not be bounded – was Van Gogh put under a time pressure to produce the painting of sunflowers?

Though it is vital that the student knows what the overall time allowance is for the test, clearly it might be helpful to the student to indicate notional time allowances for different elements of the test; if these are aligned to the relative weightings of the test (see the discussions of weighting and scoring) they enable a student to decide where to place emphasis in the test – she may want to concentrate her time on the high-scoring elements of the test. Further, if the items of the test have exact time allowances, this enables a degree of standardization to be built into the test, and this may be useful if the results are going to be used to compare individuals or groups.

Plan the scoring of the test

The awarding of scores for different items of the test is a clear indication of the relative significance of each item – the weightings of each item are addressed in their scoring. It is important to ensure that easier parts of the test attract fewer marks than the more difficult parts, otherwise a student's results might be artificially inflated by answering many easy questions and fewer more difficult questions (Gronlund and Linn, 1990). Additionally, there are several attractions to making the scoring of tests as detailed and specific as possible (Cresswell and Houston, 1991; Gipps, 1994; Aiken, 2003; Izard, 2005), awarding specific points for each item and sub-item, for example:

- it enables partial completion of the task to be recognized – students gain marks in proportion to how much of the task they have completed successfully (an important feature of domain-referencing);
- it enables a student to compensate for doing badly in some parts of a test by doing well in other parts of the test;
- it enables weightings to be made explicit to the students;
- it enables the rewards for successful completion of parts of a test to reflect considerations such as the length of the item, the time required to complete it, its level of difficulty, its level of importance;
- it facilitates moderation because it is clear and specific;
- it enables comparisons to be made across groups by item;
- it enables reliability indices to be calculated (see discussions of reliability);
- scores can be aggregated and converted into grades straightforwardly.

Ebel (1979) argues that the more marks that are available to indicate different levels of achievement (e.g. for the awarding of grades), the greater the reliability of the grades will be, though clearly this could make the test longer. Scoring will also need to be prepared to handle issues of poor spelling, grammar and punctuation; is it to be penalized, and how will consistency be assured here? Further, how will issues of omission be treated, for example, if a student omits the units of measurement (miles per hour, dollars or pounds, metres or centimetres)?

Related to the scoring of the test is the issue of reporting the results. Results may be reported item by item, section by section, or by the whole test. This degree of flexibility might be useful for the researcher, as it will enable particular strengths and weaknesses in groups of students to be exposed.

The desirability of some of the above points is open to question. For example, it could be argued that the strength of criterion-referencing is precisely its specificity, and that to aggregate data (e.g. to assign grades) is to lose the very purpose of the criterion-referencing (Gipps, 1994, p. 85). For example, if I am awarded a grade E for spelling in English, and a grade A for imaginative writing, this could be aggregated into a C grade as an overall grade of my English-language competence, but what does this C grade mean? It is meaningless, it has no frame of reference or clear criteria, it loses the useful specificity of the A and E grades, it is a compromise that actually tells us nothing. Further, aggregating such grades assumes equal levels of difficulty of all items.

If a test is designed to assess 'mastery' of a subject, then the researcher is faced with the issue of deciding what constitutes 'mastery' – is it an absolute (i.e. very high score) or are there gradations, and if the latter, then where do these gradations fall (or is it a pass/fail: either a driver can reverse safely round a corner or he can't)? For published tests, the scoring is standardized and already made clear, as are the conversions of scores into, for example, percentiles and grades.

Underpinning the discussion of scoring is the need to make it unequivocally clear exactly what the marking criteria are: what will and will not score points. This requires a clarification of whether there is a 'checklist' of features that must be present in a student's answer.

Criterion-referenced tests will have to declare their lowest boundary: a cut-off point below which the student has been deemed to fail to meet the criteria. A compromise can be seen in those criterion-referenced tests which award different grades for different levels of performance of the same task, necessitating the clarification of different cut-off points in the examination, where compensation is possible (a fail in one area can be compensated by a high pass in another, for example, pianoforte examinations).

Ebel (1979) argues that one principle in assignation of grades is that they should represent equal intervals on the score scales. Reference is made to median scores and standard deviations, median scores because it is meaningless to assume an absolute zero on scoring, and standard deviations as the unit of convenient size for inclusion of scores for each grade (see also Cohen and Holliday, 1996). One procedure is thus:

Step 1 Calculate the median and standard deviation of the scores.
Step 2 Determine the lower score limits of the mark intervals using the median and the standard deviation as the unit of size for each grade.

However, the issue of cut-off scores is complicated by the fact that they may vary according to the different purposes and uses of scores (e.g. for diagnosis, for certification, for selection, for programme evaluation), as these purposes will affect the number of cut-off points and grades and the precision of detail required. For a full analysis of determining cut-off grades, see Linn (1993).

The issue of scoring covers a range of factors, such as grade norms, age norms, percentile norms and standard score norms, for example, z-scores and T-scores (see Chapter 42), stanine scores, percentiles (see Chapter 40). Readers are referred to Cronbach (1970), Gronlund and Linn (1990), Cohen and Holliday (1996) and Hopkins *et al*. (1996) for further discussion of these.

Pilot the test

Piloting can be done in several ways:

- a small group of experts can examine the items in the test for their suitability, validity, relevance, possible cultural biases and sources of invalidity and unreliability, remoteness from the test-takers' experiences;
- a small group of test-takers, asking them to give feedback on:
  - the clarity of the items, instructions and layout
  - ambiguities or difficulties in wording
  - readability levels and language problems for the target audience
  - the *type* of question and its format (e.g. rating scale, multiple choice, open, closed etc.)
  - response categories for closed questions and multiple-choice items, and for the appropriateness of specific questions or stems of questions
  - omissions, redundant and irrelevant items
  - the clarity of the layout of the test
  - the time taken to complete the test
  - the complexity of the test items;
- involve a larger group of test-takers, to gather sufficiently large-scale data to calculate reliability levels (alphas), item difficulty and item discriminability, to identify commonly misunderstood or non-completed items, to check which items are consistently omitted or not reached (i.e. if the time was too short so that test-takers run out of time) and to be able to test out the marking scheme.

27.8 Software for preparation of a test

There are very many websites that researchers can visit to download software either free or for inexpensive purchase for test preparation, construction, layout, marking and for collation and weighting of marks, and we list these on the accompanying website. Such software does not exonerate the researcher/test deviser from the thinking that goes into the test construction; rather, it follows from that thinking and preparation, and turns it into practical formats for administration either in hard copy or online. These kinds of software packages do not address validity and reliability, and the researcher will need to pilot and refine the test before final use.

The use of software and online testing can remove some of the burden of marking, data entry and analysis, as online tests can perform these calculations automatically (e.g. for closed/multiple-choice items), and optical mark scanners can also be used to read in marks from hard copy into a computer file.

27.9 Devising a pre-test and post-test

The construction and administration of tests is an essential part of the experimental model of research, where a pre-test and a post-test must be devised for the control and experimental groups. The pre-test and post-test must adhere to several guidelines:

- The pre-test may have questions which differ in form or wording from the post-test, though the two tests must test the same content, i.e. they will be alternate forms of a test for the same groups.
- The pre-test must be the same for the control and experimental groups.
- The post-test must be the same for both groups.
- Care must be taken in the construction of a post-test to avoid making the test easier to complete by one group than another.
- The level of difficulty must be the same in both tests.

Test data feature centrally in the experimental model of research; additionally, they may feature as part of a questionnaire, interview and documentary material.

27.10 Ethical issues in testing

A major source of unreliability of test data derives from the extent and ways in which students have been prepared for the test. These can be located on a continuum from direct and specific preparation, through indirect and general preparation, to no preparation at all. With the growing demand for test data (e.g. for selection, certification, grading, employment, tracking, entry to higher education, accountability, judging schools and teachers) there is a perhaps understandable pressure to prepare students for tests. This is the 'high-stakes' aspect of testing, where much hinges on the test results. At one level this can be seen in the backwash effect of examinations on curricula and syllabuses; at another level it can lead to the direct preparation of students for specific examinations (Zhao, 2014). Preparation can take many forms (Mehrens and Kaminski, 1989; Gipps, 1994; Zhao, 2014):

- ensuring coverage, among other programme contents and objectives, of the objectives and programme that will be tested;
- restricting the coverage of the programme content and objectives only to those that will be tested;
- preparing students with 'exam technique';
- practising with past/similar papers;
- directly matching the teaching to specific test items, where each piece of teaching and contents is the same as each test item;
- practising on an exactly parallel form of the test;
- telling students in advance what will appear on the test;
- practising on, and preparation of, the identical test itself (e.g. giving out test papers in advance) without teacher input;
- practising on, and preparation of, the identical test itself (e.g. giving out the test papers in advance), with the teacher working through the items, maybe providing sample answers.

How ethical it is to undertake the final four of these, or indeed any apart from the first on the list, is questionable. Are the items cheating or legitimate test preparation? Should one teach to a test; is not to do so a dereliction of duty (e.g. in criterion- and domain-referenced tests) or giving students an unfair advantage and thus reducing the reliability of the test as a true and fair measure of ability or achievement? In high-stakes testing (e.g. for public accountability, to compare schools and teachers, for entrance to higher education and employment) there is even the issue of not entering for tests students whose performance will be low (e.g. Haladyna *et al*., 1991). There is a risk of a correlation between the 'stakes' and the degree of unethical practice: the greater the stakes, the greater the incidence of unethical practice. Unethical practice occurs where scores are inflated but reliable inference on performance or achievement is not, and where different groups of students are prepared differentially for tests, i.e. giving some students an unfair advantage over others. To overcome such problems, it is ethical and legitimate for teachers to teach to a broader domain than the test, teachers should not teach directly to the test, and the situation should only be that better instruction rather than test preparation is acceptable (Cunningham, 1998).

One can add to this list of considerations (Cronbach, 1970; Hanna, 1993; Cunningham, 1998) the view that:

- tests must be valid and reliable (see Chapter 14);
- the administration, marking and use of the test should only be undertaken by suitably competent/qualified people (i.e. people and projects should be vetted);
- access to test materials should be controlled, for instance: test items should not be reproduced apart from selections in professional publications; the tests should only be released to suitably qualified professionals in connection with specific professionally acceptable projects;
- tests should benefit the testee (beneficence);
- clear marking and grading protocols should exist (the issue of transparency is discussed in Chapter 7);
- test results are only reported in a way that cannot be misinterpreted;
- the privacy and dignity of individuals should be respected (e.g. confidentiality, anonymity, non-traceability);
- individuals should not be harmed by the test or its results (non-maleficence);
- informed consent to participate in the test should be sought.

Whilst the use of tests in research is bound by the same ethical requirements as other forms of data collection (e.g. informed consent, non-maleficence, anonymity and confidentiality, rights to non-participation and withdrawal etc.), a further major ethical issue concerns the use made of the test data (consequential validity). Here the test data should only be used for the purpose for which the test was constructed; too often test data become used for purposes other than these, and this is ethically highly questionable.

27.11 Computerized adaptive testing

Computerized adaptive testing (Wainer and Dorans, 2000; Aiken, 2003; Wainer, 2015) focuses on which particular test items to give to participants, based on their responses to previous items. It is particularly useful for large-scale testing, where a wide range of ability can be expected. Here a test is devised that enables the tester to cover this wide range of ability, hence it must include some easy to some difficult items; too easy and it does not enable a range of high ability to be charted (testees simply getting all the answers right); too difficult and it does not enable a range of low ability to be charted (testees simply getting all the answers wrong). We find out very little about a testee if we ask a battery of questions which are too easy or too difficult. Further, it is more efficient and reliable if a test can avoid the problem for high-ability testees of having to work through a mass of easy items in order to reach the more difficult items and for low-ability testees of having to try to guess the answers to more difficult items. Hence it is useful to have a test that is flexible and that can be adapted to the testees. For example, if a testee found an item too hard the next item could adapt to this and be easier, and, conversely, if a testee was successful on an item the next item could be harder.

Wainer (2015) indicates that in an adaptive test the first item is pitched in the middle of the assumed ability range; if the testee answers it correctly then it is followed by a more difficult item, and if the testee answers it incorrectly then it is followed by an easier item. Computers provide an ideal opportunity to address the flexibility, discriminability and efficiency of testing. Aiken (2003, p. 51) suggests that computer adaptive testing can reduce the number of test items present to around 50 per cent of those used in conventional tests. Testees can work at their own pace, they need not be discouraged but can be challenged, the test is scored instantly to provide feedback to the testee, a greater range of items can be included in the test and a greater degree of precision and reliability of measurement can be achieved; indeed test security can be increased and the problem of understanding answer sheets is avoided.

Computer adaptive testing has several putative attractions. On the other hand, it requires different skills from those in traditional tests, which might compromise the reliability of the test, for example:

- the mental processes required to work with a computer screen and computer program differ from those required for a pen-and-paper test;
- motivation and anxiety levels increase or decrease when testees work with computers;
- the physical environment might exert a significant difference, for example, lighting, glare from the screen, loading and running the software;
- reliability shifts from an index of the variability of the test to an index of the standard error of the testee's performance. The usual formula for calculating standard error assumes that error variance is the same for all scores, whereas in Item Response Theory it is assumed that error variance depends on each testee's ability – the conventional statistic of error variance calculates a single average variance of summed scores, whereas in Item Response Theory this is at best very crude, and at worst misleading as variation is a function of ability rather than test variation and cannot fairly be summed (see Thissen (1990) for an analysis of how to address this issue);
- having so many test items increases the chance of including poor items.

Computer adaptive testing requires a large item pool for each area of content domain to be developed (Wainer, 2015), with sufficient numbers, variety and spread of difficulty. All items must measure a single aptitude or dimension, and the items must be independent of each other, i.e. a person's response to an item should not depend on that person's response to another item. The items have to be pre-tested and validated, their difficulty and discriminability calculated, the effect of distractors reduced, the capability of the test to address unidimensionality and/or multi-dimensionality to be clarified, and the rules for selecting items to be enacted.

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Using secondary data in educational research

CHAPTER 28

This chapter introduces researchers to the use of secondary data in educational research. It raises considerations such as:

- defining secondary data
- advantages of using secondary data
- challenges in using secondary data
- ethical issues in using secondary data
- examples of secondary data analysis
- working with secondary data

28.1 Introduction

Secondary data are a valuable source for researchers, yet, despite the massive amount of such data on the Internet, they are often under-used by educational researchers (Smith, 2011). They have considerable potential for yielding important insights and foci for research (Heaton, 2008). Secondary analysis can be used to test hypotheses, to generate new knowledge and to support, challenge and extend existing theories or findings (Heaton, 1998, 2008).

Defining secondary data is not straightforward, as there are many definitions. However, generally speaking, secondary data and its analysis work on data that were originally collected for a different purpose (Glaser, 1963, p. 11) or use pre-existing data, sometimes from the same researcher but usually collected by someone else, for answering new or additional research questions or 'to pursue a research interest that is distinct from that of the original research' (Heaton, 1998, p. 1), addressing new or additional purposes, or re-analysing existing data from a new angle or with new analytical tools (cf. Vartanian, 2011). In a sense they can be regarded as second-hand data, having already been used previously. Such data often come in the form of survey data, and comprise, for example:

- official statistics;
- national surveys (census and survey data from governments or organizations), for example, the General Household Survey;
- universities' and other institutions' records and administrative data;
- international surveys and assessments, for example, Trends in International Mathematics and Science Study (TIMMS), the Programme for International Student Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS);
- ongoing databases (e.g. the National Pupil Database);
- longitudinal, regular and cohort studies (e.g. the British Household Panel Survey, the British Cohort Study, national birth cohort studies such as the Millennium Cohort Study, the National Child Development Study, the Longitudinal Study of Young People in England, the German National Educational Panel Study);
- data archives (e.g. the Consortium for European Social Science Data Archives; the European Social Survey);
- large-scale, specific surveys (e.g. the Youth Cohort Study, the Young People's Social Attitudes Survey);
- learning analytics;
- library records;
- accounts;
- administrative records (e.g. from governments and professions, such as the Department for Education).

Secondary data, both quantitative and qualitative, can also include meta-analyses, literature, reports, technical reports and summaries, scholarly journals and books, though some might argue that these are tertiary data, for example, summaries of previous secondary data.

Data for secondary analysis are held in national and international databases, for example:

- UNESCO
- OECD
- Programme for International Student Assessment (PISA)
- Trends in International Mathematics and Science Study (TIMMS)
- World Bank
- Higher Education Statistics Agency
- Higher Education Funding Council for England (HEFCE)
- government offices

- Office for National Statistics
- National Center for Educational Statistics
- data archives
- National Pupil Database
- What Works Clearinghouse
- Social Trends Database
- government department databases (e.g. the Department for Education)
- Skills Funding Agency
- Education Funding Agency
- UK Data Service
- British Social Attitudes Survey
- Universities Central Admissions Service (UCAS)
- organizations which provide information about and access to databases on educational topics
- other data providers: for example, the Inter-University Consortium for Political and Social Research.

The websites for these are provided in the companion website.

Vignoles (2007) notes that many governments are linking data from different studies by giving a unique identifier to each individual. She comments on potential benefits here but also potential ethical problems, for example, who has oversight and access and how confidentiality is to be addressed. Mostafa (2016) reports steps that have been taken to link education, health and economic records, and issues of privacy, confidentiality and consent that this raises.

Secondary data can be quantitative and/or qualitative. Quantitative data often feature in surveys, assessment data, census returns and government data sets, and can contribute to the 'political arithmetic' tradition (Smith, 2008). Heaton (2008) identifies five types of secondary analysis of qualitative data:

- supplementary analysis: in-depth analysis and additional subset/sub-sample analysis of an issue or aspect of the data that only emerged from, or which was only partially addressed in, the original research;
- supra analysis: where 'the aims and focus of the secondary study transcend those of the original research' (p. 35);
- re-analysis: to validate and confirm findings of the original study;
- amplified analysis: to conduct comparative or combined analysis of two or more qualitative data sets;
- assorted analysis: to use existing data with new primary data in the same study.

Heaton's classification alerts the researcher to the need to distinguish between microdata, metadata and results.

28.2 Advantages of using secondary data

Using secondary data has many attractions (e.g. Vignoles, 2007; Smith, 2008, 2011, 2012; Long-Sutehall *et al*., 2010; Mueller and Hart, 2010; Yorke, 2011; Gorard, 2013; Johnston, 2014; Morrow *et al*., 2014; Mostafa, 2016). For example, the scale, scope and amount of the data are usually much larger and more representative than a single researcher could gather, and the large scale and scope of data can be analysed at a level of complexity not available to smaller-scale research. Large-scale data may be more robust than small-scale data, with greater validity, and the quality of the data might be higher and more rigorous than those collected by individuals. Indeed the data have already been collected; they already exist, so the researcher does not need to worry too much about challenges of data collection, for example, financing the data collection, time taken to collect data (particularly from large samples), access to people, permissions from gatekeepers etc. Secondary data are low-cost, even free of charge, convenient and not beyond the scope of the individual researcher; they save time and money, and they are often cheaper than new research. Access to secondary data is often quick, even immediate, and this holds the advantage of timeliness and speed for decision making, in contrast to some data which require a longer time to collect; they are accessible to all and typically without much bureaucratic procedure.

Secondary data provide materials for useful descriptive analysis and the range of topics available is vast. Data come from different sources and can be combined to yield a robust analysis; they can be analysed using different approaches and perspectives which were not undertaken in the original research. Alternative relationships between variables can be explored, and new modelling techniques and statistics that were not available when the original data were produced can be tried and tested. Indeed the data may be used for research training and practice purposes, i.e. for capacity building.

As secondary data are already in existence, the research is unobtrusive and can respect ethical issues of privacy, confidentiality, anonymity, non-traceability and leaving people alone. Data may exist on sensitive topics and from 'hard to reach' people (Smith, 2008, p. 40) or an 'elusive population' (Fielding, 2004), which a researcher might otherwise have difficulty in gathering, from groups and topics to which access may be restricted or banned, for example, prisoners, victims of child abuse. Oversampling of some hard-to-reach

groups may give greater precision to the data, and data from sub-groups and minority and vulnerable groups may be sufficient for robust data analysis. Further, the data may be about a population from which a sample can be drawn.

With regard to populations and sampling, many of the data types (e.g. administrative data, government-funded research) are at the individual level, thereby enabling detailed analysis over time and contexts, and enabling systemic or ecological factors to be explored. The data may be about a population from which a sample can be drawn.

Making secondary data available serves scientific, utilitarian, pragmatic and moral arguments (for the public good and for greater benefit from public money spent on research). The data may be used for triangulation (e.g. of perspectives, sources, data, methods, time, location etc.), for longitudinal studies, replication studies, for re-analysis and re-interpretation of existing studies, and for trend analysis. They can be used to identify problems or areas for further research, to develop research questions, and for setting a context or background for further in-depth or mixed methods study, or, indeed, be a part of a mixed methods study.

28.3 Challenges in using secondary data

Secondary data are not without their challenges (e.g. Dale *et al*., 1998; Coyer and Gallo, 2005; Croxford, 2006; Vignoles, 2007; Smith, 2008; Yorke, 2011; Mostafa, 2016). For example, such data have been collected for purposes, interests, and to answer research questions and in contexts (however defined) that differ substantially from those of the present researcher, and this may bias the present research. The data may not be a sufficiently close fit to the conceptual framework, purposes, sampling or data types sought for the present research. The definitions used in the original data may be a poor fit to the present research and may change over time, and, indeed, there may be limited or no evidence on how the original data were collected, from whom and by whom, and with what response rates. Knowledge of the original research design, instruments and methods for the data collection may be unavailable and there may be limited accompanying information on the studies.

For example, in longitudinal studies, even for the same named survey (e.g. the Youth Cohort Surveys), at each time point: the designs may change; the impact of competitive tendering may affect the study (e.g. different people or institutions conduct the study); there may be inadequate, ambiguous, inconsistent or different questions and coding of responses; classifications and categories, and what is entered into them, may change over time (Uprichard, 2012); definitions and scope of issues may change; there may be a lack of continuity in content, questions, foci and analysis; there may be inadequate documentation of procedures used from the previous time; and there may be sample attrition (Croxford, 2006).

Further, there may be restrictions on how the data may be used and shared, including attention to ethical issues of informed consent and confidentiality in an era of freedom of information, and restrictions on when the data become available for public access and use. Some holders of primary research data may not permit access or may be reluctant to grant access.

Secondary data may not be neutral, but may emanate from governments, associations and institutions, i.e. those with power and with a particular agenda at the time (e.g. unemployment and welfare data and how they are measured, recognizing that how they are measured changes over time). Such data, particularly official statistics, may be social, ideological and political products (Smith, 2008, p. 79). They may be imperfect representations of the real situation, may not be very detailed or rich, or may be out of date and ill-suited to the present situation. Indeed Smith notes that the data are socially constructed, i.e. the concepts and categories they use are social constructions, and the scales for entering data are a social construction (e.g. what is meant by 'school exclusion'). The concepts may be too complex to fairly reduce to numbers, and may not be truly objective. She notes (pp. 27–9) that official statistics on items such as suicide and school exclusion may be unreliable (e.g. suicides may not be reported as such) and their measurement may be fallible. This does not necessarily mean that they are not valuable, but that researchers have to be cautious in using them (p. 29).

There are other practical issues in using secondary data, for example: the data may be ambiguous and contain errors (which might be impossible to know) and the data may have been saved selectively, i.e. some data may be excluded from the original study; some constructs or composite factors may be defined operationally by a limited range of variables. The data might not address the important issue of *why* something is as it is or was as it was; i.e. they offer descriptions rather than explanations (this may not be a problem, indeed Campbell *et al*. (1982) see this descriptive function as a powerful and important stage in research, or even as the main topic of research itself).

Secondary data may not sustain comparative analysis, being rooted in local, regional and national cultures, socio-cultural, spatial and temporal contexts,

for example, test items (Smith (2008, p. 35) gives the example of the TIMMS study), and they may have a historical bias because of events leading to the original research.

Researchers working with secondary data must ensure that their present purposes are sufficiently compatible with the original data sets on which they are working and that the data lend themselves fairly to the kinds of analysis being used and purposes for which they are being used. This means comparing the researcher's purposes with those for which the original data were collected, and being aware of possible biases in the data. The researcher must be sensitive to the commissioners of the original data: for example, if they are governments or interest and advocacy groups, the data might be biased. This applies not only to governments but to associations, institutions and academic departments, and such information can often be found in the manuals and reports that accompany the original research.

28.4 Ethical issues in using secondary data

Simply because data are already in existence does not absolve the researchers from addressing ethical issues. Morrow *et al*. (2014) remark that, regardless of whether the data are or are not public, the researcher has to address responsibilities to the original participants, including the original researchers, and must avoid misinterpretation of the original findings and their contexts (e.g. in cross-cultural research or if researchers from one culture are using data from another, original culture).

In providing qualitative data for secondary analysis, Morrow *et al*. also note that making data available for re-use through the sharing of archived data, rather than destroying data at the end of a project, has become much more widespread with open access, and this raises ethical issues concerning both benefits and risks (see also Mostafa, 2016). On the one hand, the benefits – scientific, utilitarian, moral and pragmatic – of secondary analysis outlined earlier in this chapter are plain. On the other hand, researchers have to consider the original informed consent, confidentiality, anonymity, data protection, the avoidance of harm (including stigmatization, prejudice, misrepresentation or marginalization), ownership of data (whether it is acceptable to have such shared, common ownership, not least if the intentions of the later usage contradict those of the original use) (cf. Mauthner, 2012; Mostafa, 2016), and the ethics of using data in ways for which they were not originally collected or intended.

Researchers and participants in the original study will not know how information which is archived will be used; as Morrow *et al*. (2014) remark, when researchers seek the consent of participants, in effect they are asking them to consent to something which is uncertain, as the future and the future use are uncertain, thereby rendering informed consent as not really being fully informed (p. 10). It may be possible, for example, to identify participants by the quality and rich detail that they provide, even with anonymization and anonymity; here secondary analysis faces a challenge of how much data to report, as rich, context-laden data – the stuff of qualitative research – may breach anonymity and non-traceability.

Addressing anonymity and non-traceability, avoiding misrepresentation and abuse of data, doing no harm, just as in other forms of research (see Chapter 7), should prevail. Secondary data researchers must consider their responsibilities to the original participants.

28.5 Examples of secondary data analysis

Examples of secondary data analysis are plentiful (cf. Heaton, 2008; Smith, 2008). For example, Smith (2012) looks at UK National Census data in relation to the Sure Start programme and at UCAS data on recruitment to higher education. Her earlier volume (2008) contains very many worked examples of different studies in a range of fields of education, using data from different education sources.

Rutkowski *et al*. (2010) comment on how to improve secondary analysis and reporting for international large-scale assessment data, and Hampden-Thompson *et al*. (2011) report an example of how large-scale secondary data analysis can be combined with in-depth qualitative approaches. Mueller and Hart (2010) report an example of secondary analysis of archived data on gifted education, and comment on two major sources of archived data sets (the National Longitudinal Study of Adolescent Health and the Educational Longitudinal Study). Homer *et al*. (2011) use the National Pupil Database to investigate performance across a range of science courses, using different valued-added approaches. Their study draws attention to problems for research deriving from students changing schools (7.7 per cent of students between Key Stages 3 and 4) and missing data (some 6.6 per cent of cases).

28.6 Working with secondary data

In working with secondary data the researcher must understand several matters concerning the original

research in relation to the present research (Smith, 2008; Newby, 2010; Johnston, 2014), including:

- the original purposes, conceptual and theoretical structure of the data (Smith, 2008, p. 62), and the inclusion and exclusion of variables and indicators in the original research that might be important for the present researcher, i.e. a comparative analysis of necessary and important variables and indicators (p. 65), and the relevance and suitability of the data for the researcher's present purposes;
- the definitions of terms being used in the original and present research. This is a particular problem with definitions emanating from those in power, for example, governments, who may have definitions which do not accord with the researcher's or which may change markedly over time, for example, definitions of 'employed' and 'unemployed';
- the type and quality of the original data, for example, were they 'hard' data, perception-based or opinion-based data, and how was verification undertaken of 'hard' data (and how they were collected);
- who collected the data (e.g. trained researchers and data collectors, data-collection agencies, market researchers, opinion pollsters) and the credentials and sources of the authors of the original data;
- who gave the original data (i.e. how trustworthy are the original data), how were the original data collected (e.g. online, face-to-face, by post, by telephone, from professional registers and records) and what incentives, if any, were used to encourage participation and response?
- what was the sampling strategy in the original research, and how does it fit the present researcher's purposes; what were the response rates and attrition rates, and which groups were over-represented, under-represented or absent, i.e. sampling bias and population characteristics (such data are usually present in the research manuals and reports for the original research)?
- what questions and kinds of questions were asked and how were they asked, what response categories and scales were used (e.g. Likert scales, rankings) in the original data and how compatible are they to the present researcher's purposes (i.e. are the original instruments for data collection available so that the researcher can check for compatibility and utility with the present research purposes and contents)?
- the date of publication of the original data (i.e. are the data now out of date?), their intended audience, the intended and actual coverage of the topic or issue, in what socio-temporal contexts, and how well these match the present researcher's contexts, situation and purposes (i.e. a question of relevance between the original and the contemporary context);
- how factually correct were the original data; what and how much missing data were there in the original data set?
- the format of the data archive (the researcher may need to rework the format, which can be time-consuming);
- how were the original data analysed and grouped (which could affect how they were presented in the original data set); what level of analysis was used (e.g. individual, organizational, local, regional, national); what was the unit and scale of data collection used (e.g. individual, organizational, local, regional, national etc.); how were the data summarized, presented, categorized and grouped in the original data set, how was this reached, and how compatible are they to the present researcher's categories and purposes?
- what ethical issues need to be addressed in using secondary data, for example, anonymity, informed consent (which may not be a problem as identifying data may not be included in the original data, and data may be aggregated so as to render individuals invisible or non-traceable, but may still be a problem inasmuch as the givers of the original data may not have given consent to the data being used for purposes other than those at the time).

One should not be put off by this formidable list of cautions. Rather, they direct the researcher to examine the compatibility of the original data and research with the present research, and the reliability and validity of the original research, so that its usefulness for the present research can be validated.

Researchers working with secondary data need to be clear on how they are going to analyse the data (e.g. by subject, topic or theme; by groups of people; by institution; by year; by region), and many databases enable the researcher to interrogate data by these different variables.

Yorke (2011) gives the example of using data from the National Student Survey to identify several cautions in using secondary data for research. Included in these are that the original data may not fit the kinds of data that the researcher would have preferred to collect (p. 257), and that the researcher has to decide whether the data are 'good enough' ('satisficing') for the present research, as 'good enough' for one purpose may not be so for another. He advocates a risk analysis, to judge the acceptability of using data gathered for one purpose

to serve another purpose, and from one context to another, and he notes that the researcher should make clear any compromises made and limitations encountered by changes of purpose and context.

Yorke advises that secondary data have to be checked for their quality and to avoid assuming that they are free from error (e.g. in data entry). Indeed he notes that earlier versions of data (e.g. student records) may be overwritten by later versions. Further, he notes that errors in categorization might feature in the original data (he cites the example of the Higher Education Statistics Agency in the UK in respect of degree classification), and that these errors may only be found when one is actually working with the data, i.e. when 'implausibilities become apparent' (2011, p. 259). With regard to categorization, he makes the point that the categories and categorizations used in the original data may not be compatible with contemporary categorizations or researcher's needs.

Yorke also questions just how 'national' some data are. He gives the example of trends in honours degrees in the UK's national statistics that omitted Scotland because its honours programmes differed from those in the rest of the UK, and programmes in some subjects did not use an honours classification system, i.e. a definitional, classification problem.

Yorke advises researchers to be prepared to spend time reformatting original data for analysis, and he provides examples of this, for example, how subjects are aggregated in publicly available data in the National Student Survey. This echoes Windle's (2010) comment that using secondary data analysis may not be as time-saving as researchers first imagine, as time must be spent understanding the original research, for example, its conceptual framework, research purposes and questions, contexts, design, methods, definitions, strengths and weaknesses etc. (p. 323). This is a useful caution for those researchers who might think that secondary data can save time overall or that they can be automatically used in the original format in which they were accessed. Also different levels of analysis may exist in the original data, and this may be a definitional problem.

In working with secondary data, we suggest nine practical steps that researchers can take:

Step 1: Identify your own research purposes and research questions. Search and select possible secondary data sets and data for suitability for your own research. Were you part of the original primary research?

Step 2: Familiarize yourself as much as possible with the original research and data that you have selected. This involves attention to: conceptual and theoretical frameworks; design; instrumentation; sampling and populations; variables included; definitions used; data-collection instruments and methods; completeness, accuracy and missing data (for both qualitative and quantitative data); validity and reliability of the original data set; whether the original data set included primary and/or secondary data; data organization and levels of organization; data analysis and weightings; reporting; limitations and errors.

Step 3: Assess and evaluate the data to see if they are amendable to secondary analysis and suitable for your present research: sufficient, valid, relevant, appropriate, reliable, in sufficient detail and depth, and if their scope and coverage are adequate.

Step 4: Check if there are restrictions (and copyright) on how the data may be used, analysed, disseminated, shared and published.

Step 5: Check that ethical issues have been addressed (e.g. what consent was given in the original study, how to address non-traceability and anonymity, how to address any conditions required in the original study).

Step 6: Check the fit and congruence between the original study and your own study: (a) its conceptual and theoretical framework; (b) its socio-cultural, temporal, locational and political contexts; (c) its research purposes, design and research questions; (d) the variables that it included; (e) its population and sampling; (e) the definitions and categories that it used; (f) its data types and coverage; (g) the units of analysis and levels of analysis that it used.

Step 7: Prepare your data: reformat the data if necessary to fit your purposes; select the cases and variables of interest from the original data set; weight the cases (if desired) and recode the variables if necessary (creating new variables if necessary); decide how to take account of missing data.

Step 8: Analyse the data to address your research purposes and research questions. Analyse groups and sub-groups of cases if required.

Step 9: Decide how you will report the secondary analysis and its findings (including reference to the original study/source and its data).

Using these nine steps can guide researchers to evaluate the existing data using several criteria to ensure suitability for their present research. Secondary data are

a valuable source for researchers. As Johnston (2014) remarks, given that we live at a time of massive amounts of data being collected, compiled and archived and made accessible, the time is ripe for secondary data analysis (p. 626).

28.7 Conclusion

Secondary data are a valuable and often easily accessible resource for researchers. Whilst they have many attractions, they also bring several challenges and we advise researchers to be very aware of these and to work out how to address them. Often they cannot be used in their original form, and researchers must be prepared to put them into a format suitable for their own research and examine the sources, definitions, validity, completeness and reliability of the data etc., and take into account the contexts that gave rise to the original research and data. As with all research, using secondary data raises several ethical issues, and we advise researchers to be mindful not only of ethical issues more generally but particularly of those that feature strongly in this form of research. Fitness for purpose and match between the original research and the present research should be sufficiently close to render the use of secondary data valid.

Researchers working with secondary data might need to consider: costs; feasibility; meta-analysis (if original microdata are available); and replication. Readers may also find it helpful to go to Chapter 16 on historical and documentary research.

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Personal constructs

CHAPTER 29

Richard Bell

This chapter discusses the following:

- strengths of repertory grid technique
- working with personal constructs
- grid analysis
- examples of the use of the repertory grid in educational research
- competing demands in the use of the repertory grid technique in research
- resources

29.1 Introduction

Personal constructs are the basic units of analysis in a complete and formally stated theory of personality proposed by George Kelly in his book *The Psychology of Personal Constructs* (1955). Kelly's own clinical experiences led him to the view that there is no objective, absolute truth and that events are meaningful only in relation to the ways that are construed by individuals. Kelly's primary focus is on the way individuals perceive their environment, the way they interpret what they perceive in terms of their existing mental structure and the way in which, as a consequence, they behave towards it. In *The Psychology of Personal Constructs*, Kelly proposes a view of people actively engaged in making sense of and extending their experience of the world. Personal constructs are the dimensions that we use to conceptualize aspects of our day-to-day world, and, as Kelly writes, people differ from each other in their construction of events. The constructs that we create are used by us to forecast events and rehearse situations before their actual occurrence, and are sometimes organized into groups which embody subordinate and superordinate relationships. According to Kelly, we take on the role of scientist seeking to predict and control the course of events in which we are caught up. For Kelly, the ultimate explanation of human behaviour 'lies in scanning man's [*sic*] undertakings, the questions he asks, the lines of inquiry he initiates and the strategies he employs' (Kelly, 1969, p. 69). Education, in Kelly's view, is necessarily experimental. Its ultimate goal is individual fulfilment and the maximizing of individual potential, capitalizing on the need of each individual to question and explore.

Kelly's theory was very formally constructed, with a 'fundamental postulate' and eleven corollaries that followed from this. Later, even in Volume II of this work, but more obviously even later (Kelly, 1969), he moved away from this very formal statement. Butt (2008) provides a good introduction to this broader view of Kelly and his theory. Nevertheless, the formal statement provides the model for the repertory grid. The fundamental postulate is:

> *A person's processes are psychologically channelized by the ways in which he or she anticipates events.*

One key component in the above statement is *the ways in which*. For Kelly, these *ways* are called constructs. The person's repertoire of constructs is the basis by which the person construes or understands their world and makes predictions about the future. What is to be construed or understood are the *events*. What are these events? It is rather an ambiguous term. In one sense they really are events such as 'going to a party' or 'teaching a class', but Kelly uses the term much more broadly to encompass all psychological objects such as 'ideal self' or 'teacher I looked up to'. In the repertory grid technique, these objects are termed *elements*.

A repertory grid then is simply a representation of the relationship between these elements and constructs. As such it provides information we can use to understand *how* 'a person's processes are psychologically channelized by the ways in which he or she anticipates events' (Kelly, 1955, p. 46). Figure 29.1 shows a simple grid layout for collecting data.

Kelly's theory then has a number of corollaries to this fundamental postulate that relate to the constructs. An important one from a repertory grid perspective is that constructs are essentially bipolar, that is, capable of being defined in terms of polar adjectives (good–bad) or polar phrases (makes me feel happy–makes me feel sad). Other corollaries can also affect the ways in

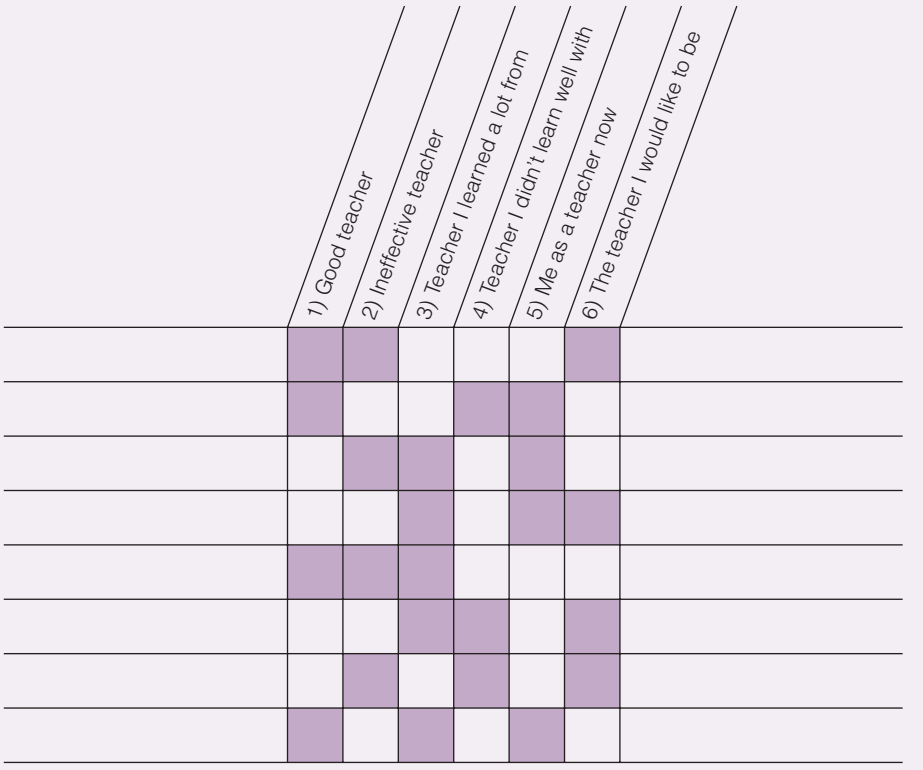


FIGURE 29.1 Simple grid layout

Instructions: Consider the three figures represented by the shaded cells in each row. Which two are more alike and by the same token different from the third? Enter the quality of the similar pair on the *left*-hand side of that row, and the quality that differentiates the third person on the *right*-hand side.

which we can use the repertory grid technique, as we shall see later. In addition to the corollaries there are other formal aspects of the theory that play roles in relating constructs to behaviour. Fransella (2003, pp. 455–7) provides a concise summary of all components of the theory.

A number of different forms of repertory grid technique have been developed since Kelly's first formulation. They have the two essential characteristics in common that we have already identified, that is, constructs – the dimensions used by a person in conceptualizing aspects of his or her world – and elements – the stimulus objects that a person evaluates in terms of the constructs she employs. In Figure 29.1, we illustrate the empirical technique suggested by Kelly for eliciting constructs and identifying their relationship with elements in the form of a repertory grid.

Since Kelly's (1955) original account of what he called 'The Role Construct Repertory Grid Test', several variations of repertory grid have been developed and used in many different areas of research. In a chapter entitled 'Some Uses to Which Grids Have Been Put', Fransella *et al*. (2004) provide an annotated bibliography of a wide range of areas including 'working with children', 'teachers and teaching', 'construing of professionals' and 'those with learning difficulties'.

29.2 Strengths of repertory grid technique

The repertory grid technique draws its strength from two particular features. The obvious one is that it is an individualized technique where the respondent provides the framework as well as the responses. For example,

Suto and Nádas (2009) used grids with two principal examiners to identify features of examination questions that differed in the difficulty of marking. The other, perhaps more important, strength derives from the two-way nature of the data where elements are related to constructs. This enables relationships between elements to be assessed, since there is information about each element provided by the set of constructs. Conversely, the relationships between constructs can be examined through the information provided for each construct by the set of elements. Even if the individuality of the grid is restricted by the use of provided constructs or elements (as discussed below), the two-way data allows for within-respondent analyses to be carried out.

29.3 Working with personal constructs

Choosing elements

The key issue in choosing elements is that they should be a homogeneous set to ensure that the constructs elicited from some of them will also be relevant to other elements. Yorke (1983) and Wright and Lam (2002) draw attention to problems that can arise when this requirement is not met.

In Kelly's original technique, elements were usually chosen by the client to fit 'role titles' provided by the clinician. Some role titles allowed no choice ('me as I am now'); some allowed a possible choice ('your mother (or the person who filled that role in your life)'); and some allowed wide choice ('a teacher you admired'). Some research (Bell *et al*., 2002; Haritos *et al*., 2004) suggests that value-laden role titles such as the teacher role above or 'a girl you did not like' tend to polarize the grid by making constructs subsequently elicited more similar to one another than when role titles are neutral, such as 'a significant person in your life'.

Eliciting constructs

Kelly originally suggested six ways in which constructs could be elicited from elements, the most familiar being to choose three elements and ask the participant to specify *some important way in which two of them are alike and thereby different from the third*. The way in which two were alike formed one pole (the similarity pole), the way in which the third differed formed the other pole. Another way involves asking the participant for the opposite to the similarity pole. Another way suggested by Kelly was to always have the 'as I am now' element always included. This is more widely used in psychotherapy settings since it ensures that all constructs are relevant to the self.

The task of triadic comparison is cognitively demanding at it has been found that simply using two elements is better for children (e.g. Salmon, 1976) or those with learning difficulties (Barton *et al*., 1976).

'Elicited' versus 'provided' constructs

One form of repertory grid technique now in common use represents a significant departure from Kelly's original procedure in that they provide constructs to subjects rather than elicit constructs from them. Eliciting constructs from individuals follows from Kelly's *individuality corollary*: *Persons differ from each other in their construction of events*. Supplying constructs contravenes this, though Kelly also posited a *commonality corollary*: *To the extent that one person employs a construction of experience which is similar to that employed by another, his or her psychological processes are similar to those of the other person*. This suggests that there will be constructs which are common to a number of individuals.

Can the practice of providing constructs to subjects be reconciled with the individuality corollary assumptions? Despite much research, the answer is still unclear and of course is further clouded by the more recently discovered effect of value-laden role titles in elicited constructs. As Fransella *et al*. (2004, p. 48) point out, however, '[c]onstructs have to be supplied in a group context if group data is required'. Bell (2000) has shown how the commonality corollary may be simply tested by examining each supplied construct in turn for unidimensionality of the element ratings.

But the issue of supplied or elicited constructs is not necessarily an all-or-none situation. Bannister and Mair (1968) support the use of supplied constructs in experiments where hypotheses have been formulated and in those involving group comparisons. The use of elicited constructs alongside supplied ones can serve as a useful check on the meaningfulness of those that are provided, substantially lower inter-correlations between elicited and supplied constructs suggesting, perhaps, the lack of relevance of those provided by the researcher.

Allocating elements to constructs

In Kelly's original technique participants were allowed to classify as many or as few elements at the similarity or the contrast pole, giving a very lopsided construct. Originally this was seen as a problem since measures of association between constructs could be affected by this (e.g. Bannister and Mair, 1968, p. 59) and strategies were proposed to overcome it. These strategies had problems of their own, in that they forced the participant to allocate elements to constructs in a fixed fashion, as removed from Kelly's individual focus as

are supplied constructs. More recently Bell (2004a) has shown that lopsidedness is associated with the super- and subordinate relationships implied by Kelly's organizational corollary. The common method now of allotting elements is the 'rating form', which tends to lessen the tendency to locate elements on a construct in a lopsided fashion. Here, the subject is required to judge each element on a multi-point scale, where one extreme (say, 7) is aligned with one pole ('notices when I am having problems') and the other extreme (1) is aligned with the other pole ('doesn't notice when I am having problems'). The rating form is illustrated in Figure 29.2.

As with most rating scale formats, questions arise as to the meaning of a mid-point rating (when an odd number of rating points are specified). Another of Kelly's corollaries was the range corollary: *a construct is convenient for the anticipation of a finite range of events only*. In a rated grid, then, does a mid-point rating mean neither or both poles are relevant? Or more generally, does a grid allow missing data? Kelly's range corollary would suggest 'yes' but the computation of summary measures to represent grids would usually say 'no'. In practice this problem can be ameliorated, if not overcome, by the careful nomination of a homogeneous set of elements. The mid-point issue remains intriguing. Winter *et al.* (2010) found in a psychotherapy study that constructs where either 'self now' or 'ideal self' was located at the mid-point were associated with more complex cognitive structures in the grid.

Other techniques: laddering and pyramid construct elicitation and other forms of grids

The technique known as 'laddering' arises out of Hinkle's (1965) linking of the notion of implication with the organization corollary: *each person characteristically evolves for his or her convenience in anticipating*

	1) Good teacher	2) Ineffective teacher	3) Teacher I learned a lot from	4) Teacher I didn't learn well with	5) Me as a teacher now	6) The teacher I would like to be	
Quiet	1	5	2	2	3	1	Loud
Sociable	4	1	5	5	1	2	Aloof
Open	2	4	1	3	1	2	Private
Creative	2	4	1	5	5	1	Follow set plans
Independent	1	5	1	3	5	1	Dependent
Listens	1	5	1	5	3	1	Doesn't listen
Reject ideas	5	1	4	1	3	5	Accepts
Strict	1	3	1	5	5	2	Lax

FIGURE 29.2 Completed grid

Instructions: Consider how the qualities in each row apply to each of the six figures. To the extent that the quality on the *left* applies more to the person, give a rating closer to *1*. And to the extent that the quality on the *right* applies more to the person, give a rating closer to *5*.

events, a construction system embracing ordinal relationships between constructs. Hinkle's innovation was to replace 'anticipation' with 'implication'. The linking of implication with ordinal relationships enables logical relationships to be specified between poles of different constructs. 'Laddering' is an exploratory technique using implication to move from a pole of a given construct to a pole of an as yet to be elicited construct. It usually proceeds by asking the participant to indicate which pole of the given construct is preferred. (This is linked to Kelly's *choice corollary*: *a person chooses that alternative ... through which he anticipates the greater possibility for the extension and definition of his system*.) Having identified the preferred pole, the participant is then asked 'Why?'. The response to this forms one preferred pole of the higher-order or implied superordinate construct. The construct can then be completed by asking the participant for the contrasting pole of the new construct. In turn the participant can then be asked for the preferred pole of this new construct and again asked 'Why?' to produce the first pole of the next higher-order construct. Although he never published his development of the technique of laddering, it has been used widely in many fields, particularly in research into consumer perceptions (Reynolds and Gutman (1988) provide practical advice in using this technique in the context of laddering from product properties to consumer values).

Laddering is a technique for eliciting constructs in terms of a single element, the self ('Which pole do you prefer and why?'). Pyramiding, a similar procedure developed by Landfield (1971), also uses a single element. Respondents are asked to think of a particular 'element', a person, and then to specify an attribute which is characteristic of that person. Then the respondent is asked to identify what kind of person would not have that characteristic. The researcher then returns to the first characteristic and asks 'What more can you tell me about a person who has that characteristic?' and again 'What is the opposite of that characteristic?' The enquiry is then repeated similarly for the opposite pole of the first characteristic. According to Landfield the enquiry then proceeds to similarly enquire of the four construct poles thus elicited. Landfield termed this a 'pyramid' since it starts from one element to produce two construct poles which in turn produce four construct poles and finally eight construct poles. This kind of enquiry asks for elaborations (What more can you say?) rather than implications (Why?), and thus does not identify superordinate relationships between constructs as does Hinkle's procedure.

Landfield saw his technique as purely qualitative and informing the psychotherapeutic process. Hinkle (1965), however, went on to develop an Implication Grid or Impgrid, in which the subject is required to compare each of his constructs with every other to see which implies the other. Table 29.1 illustrates Hinkle's laddering technique with an example from educational research reported by Fransella (1975).

Exchange grids are procedures developed to enhance the quality of conversational exchanges. Basically, one person's construing provides the format for an empty grid which is offered to another person for completion. The empty grid consists of the first person's verbal descriptions from which his ratings have been deleted. The second person is then invited to test his comprehending of the first person's point of view by filling in the grid as he believes the other has already completed it. Various computer programs ('Pairs', 'Cores' and 'Difference') are available with the Rep Plus package to assist analysis of the processes of negotiation elicited in exchange grids.

In the 'Pairs' analysis, all constructs in one grid are compared with all constructs in the other grid and a measure of commonality in construing is determined. 'Pairs' analysis leads on to 'Sociogrids', in which the pattern of relationships between the grids of one group can be identified. In turn, 'Sociogrids' can provide a mode grid for the whole group or a number of mode grids identifying cliques. 'Socionets' which reveal the pattern of shared construing can also be derived.

Grid administration

The way in which a grid is administered depends in part on the purpose and nature of the research. Where the researcher wants detailed information from few respondents, then administration is best carried out with one-on-one interviews. This has the advantage of allowing the researcher to monitor the constructs elicited for duplication or difficult-to-understand pole labels. Such administration usually begins with some discussion of the area the grid will relate to and a simple trial run of two or three constructs elicited from half-a-dozen elements not related to the main task. Clarifications can be sought, and the preferred pole identified as the constructs are elicited.

Grids can also be collected in group testing. This usually requires a somewhat smaller grid (the study of Haritos *et al*. (2004) used this approach) and a pro forma form that allows spaces for element and construct labels as well as grid ratings, and uses shading or some other method of identifying which elements are to form the triads (see Figure 29.2). Here overheads or PowerPoint are needed to illustrate how the grid is to be completed.

TABLE 29.1 LADDERING DIALOGUE

Okay so **good teacher** and **teacher I learned a lot from** are alike in that they are both 'alert' while **an ineffective teacher** is … *'has his mind on other things'*	The opening construct dialogue
So if you had to choose between 'having your mind on other things' or 'being alert', which would you choose *obviously I'd choose being 'alert'*	Choosing the preferred pole
Why? *well, 'being alert' means you can pick up on where each kid is at in their work but having your mind on other things means you just see the class as a whole*	Laddering up to next higher construct
Why is it important to 'pick up on where each kid is at in their work'*?* *Because kids don't all learn in the same way or at the same rate … they're…* They're individuals? *Right*	Laddering up again to next higher construct. Notice the interviewer doesn't look for the opposite pole of this construct
Why is important to treat them as individuals? *So they can each reach their potential*	Notice the interviewer doesn't look for the opposite pole of 'individuals' but keeps laddering from this pole
Why is that important? *It's important because that is why I want to be a teacher – to enable kids to realize their potential*	The interviewer stops here, sensing perhaps that this is far enough up this ladder for the moment
Can I just go back to what you said as the opposite of 'pick up on where each kid is at in their work', you said 'just see the class as a whole' *Mmm*	The interviewer now goes back to the contrast pole of the first laddered construct
What do you think seeing the class as a whole implies? *You mean about a teacher who does that?* Yes *I guess it means that they don't really care about the kids*	Laddering up to next higher construct from this contrast pole
And what would that mean? *They would be thinking about themselves – it might be temporary like some problem at home – or it might be that they want an easy life, not have to work as hard*	Laddering up to next higher construct. Notice the respondent gives two consequences
Yes, I guess there are times when temporary problems affect our work. But what about wanting an easy life – what does that imply? *Unambitious – I don't mean about promotion or things like that, but not being ambitious about being a good teacher*	The interviewer chooses one to ladder from
And the opposite of 'not being ambitious about being a good teacher' is? *Being ambitious about being a good teacher*	Here the interviewer switches attention across from this negative pole to the contrast positive pole
Is that important to you? *Yes*	And checks that it is the preferred pole
Which is more important to you: 'enabling kids to reach their potential' or 'being ambitious about being a good teacher'? *They're the same thing really; a good teacher is one who enables kids to reach their potential*	The interviewer now draws the higher-level positive poles together
And the 'being ambitious'? *Yeah, it's also important to try to improve your teaching.*	The interviewer notices that part of one of the poles has been left out and draws that in

Where grids are to be collected with supplied elements and constructs (perhaps following on from some individually elicited grids, as in Reid's and Holley's (1972) study of choice of university), the grid data can be collected in a simpler questionnaire-type format. Should the ratings be collected construct by construct, rating each element in turn, or element by element, rating each construct in turn? Evidence (Bell *et al*., 2002; Neimeyer and Hagans, 2002) suggests that it does not matter.

In this century it is to be expected that computer administration is to be a major way in which grid data is collected. A web version can be found at http://webgrid.uvic.ca for the Gaines and Shaw program Webgrid Plus that allows relatively simple grids to be elicited. Idiogrid (at www.idiogrid.com) is a freeware program that allows quite complex grid elicitations to be structured and subsequently used to collect data from respondents. Both of these resources provide for data analysis of the grids collected.

29.4 Grid analysis

Before we analyse the grid data, we need to be aware that the orientation of the data in the grid will be a function of the way in which the constructs were elicited. If the two elements that are alike are both positive figures (e.g. 'ideal self' and 'best friend') then the pole corresponding to their similarity will reflect this (e.g. generous). If the two figures are negative ones (such as 'person I dislike' and 'worst teacher') then the pole corresponding to their similarity (e.g. stingy) will reflect this – and the correlations with other constructs would be opposite.

A preliminary step in analysing a grid should often be to make all the constructs similarly aligned. This can be done by asking the respondent to indicate their 'preferred' pole. (This is also done in a related personal construct technique called laddering.) Another way of doing this is when the grid contains the element 'ideal self'. Whichever pole of each construct is aligned with the 'ideal self' is the preferred pole. Of course, this may not be of any use when the 'ideal self' is located at or near the mid-point. There is also an automatic way of doing this. We can analyse the correlations between the constructs and identify those constructs which generally correlate negatively with other constructs, and reverse these constructs.

How do we extract information from a repertory grid? There are a number of ways we can look at the numerical information in a grid, and these are discussed below.

Looking at relationships between elements and between constructs

Even in an individual grid, there is replicated information for elements (across constructs) and constructs (across elements). We can use this to make comparisons between constructs or between elements, or create indices to represent these comparisons. One of the oldest of these is Bieri's (1955) index of cognitive complexity/simplicity. This was originally a matching coefficient calculated for each pair of constructs and summed for the whole grid. These days it is usually based on the average correlation among constructs. If this is a large value, then it means all constructs are highly correlated and relate to the elements in much the same way. The person might thus be said to be construing their world in a simple way. If the average correlation is low, then the constructs are differentiated and the person might be said to be construing in a complex fashion. We can also compute such averages for each construct to determine which constructs are like the others and which are different. Figure 29.3 shows the average (root-mean-squared) correlations for each construct in the same grid. This output was obtained from the comprehensive (and easy to use) free web-based repertory grid analysis OpenRepGrid on Air package (www.onair.openrepgrid.org) (Heckmann, 2016).

Except where otherwise noted, all analyses shown in this chapter were produced with this package. In this example it can be seen that the construct '*sociable – aloof*' is least related to other constructs, while overall there is a consistent and substantial similarity in the ways these constructs are applied. Of course the construct correlations that form the basis of this index can also be analysed in other ways, such as with principal components. Figure 29.3 shows that the component loadings can be used to show where the poles are reversed with respect to the orientation of the other constructs. All construct loadings have positive signs except for '*sociable – aloof*' and '*rejects ideas – accepts*'. While the latter construct poles can easily be seen to be unaligned with the others (where the left-hand pole is the positive quality), *sociable* rather than *aloof* would normally be seen as positive. For this grid, however, *aloof* is perceived as the more positive quality in a teacher.

Cluster analysis is another way of depicting relationships in a matrix of measures of association among elements or constructs. Figure 29.4 shows element Euclidean distances and a hierarchical clustering of these. *Good teacher* and *ideal teacher [teacher I would like to be]* are similar, while *teacher I learned a lot*

```
####################################
Root-mean-square correlation of constructs
####################################

                                    RMS
(1)  quiet – loud                  0.63
(2)  social – aloof                0.32
(3)  open – private                0.48
(4)  creative – follows set plans  0.70
(5)  independent – dependent       0.73
(6)  listens - doesn't listen      0.75
(7)  rejects ideas – accepts       0.73
(8)  strict – lax                  0.62

Average of statistic 0.62
Standard deviation of statistic 0.14
```

```
###############
PCA of constructs
###############

Number of components extracted: 1
Type of rotation: varimax

Loadings:                          PC1
quiet – loud                      0.82
social – aloof                   -0.36
open – private                    0.62
creative – follows set plans      0.90
independent – dependent           0.93
listens - doesn't listen          0.96
rejects ideas – accepts          -0.93
strict – lax                      0.80
```

FIGURE 29.3 Grid summary measures

from is also similar to these two. *Me as a teacher now* is weakly associated with an *ineffective teacher* and somewhat less associated with *teacher I did not learn from*.

Particularly useful elements where grids involve the self as an element are the elements *self now* and *ideal self*. The discrepancy between these two can be taken as a measure of self-esteem or used to generate a self-identity plot (as in Figure 29.5, taken from the Idiogrid program) where self and ideal are reference axes against which are plotted the other elements.

Looking at the overall grid

In the 1960s Patrick Slater introduced a technique that he called 'principal components' and which we now know as singular-value-decomposition that enables both elements and constructs to be represented together. There have been a number of different ways of representing the constructs and elements in these maps. Elements always appear as points in the map. Constructs, however, are shown in different ways. The construct data is like a principal component solution with two columns of coordinates which define a point in the space. However, representations differ. Often the point is reflected back through the origin to make a line symmetric about the origin, the two ends of which represent the construct poles. Other representations (such as that originally used by Slater) show the construct poles as points on a circle encompassing the elements. Figure 29.6 shows a spatial representation of our sample grid with elements shown as square points and constructs shown as vectors (lines) symmetric about the origin. The longer the line, the more important the construct. The horizontal dimension separates better and poorer teachers, with constructs similarly aligned, while the vertical axis is aligned with the isolated construct *aloof – sociable* distinguishing principally between *teacher I didn't learn well with* and *me as a teacher now*.

The analyses described above all focus on variation within a grid. For many research purposes, the unit of analysis is the person (the subjects of the experiment) and consequently the focus is on between grid differences. This issue is explored later in this chapter.

29.5 Some examples of the use of the repertory grid in educational research

Jones (1999) used repertory grids alongside interviews and participant observation to elicit headteachers' views of their roles and agenda in changing times. While the study found an increase in their management activities (one construct), it also found that not only did their changing role not lead to their de-professionalization but also their core values were rooted in their values in, and views of, education (a second construct). The superordinate constructs for the primary headteachers were child-centred and management, in that order, i.e. the management systems were there to serve the

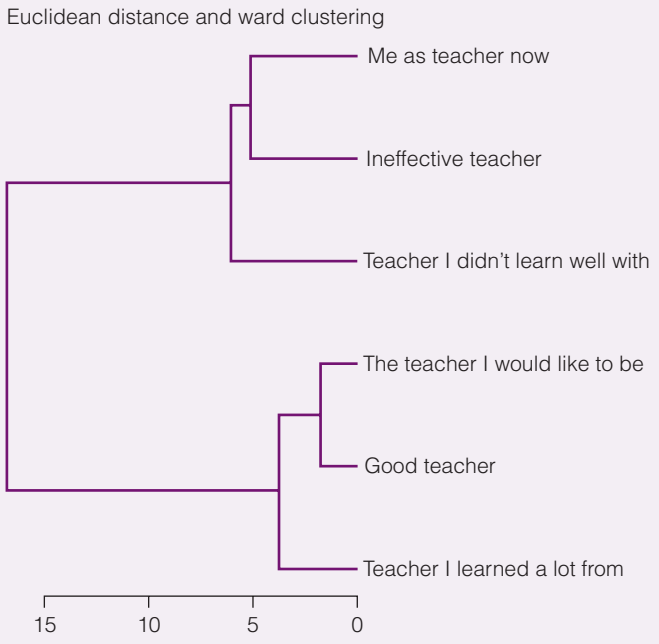


FIGURE 29.4 Grid cluster representation

child-centred values and vision. Constructs elicited included: child-centred problem solving, implementation policy, evaluation, involving other agencies, problem solving and paperwork.

Bezzi (1999) used repertory grids to explore the perceptions of the images of the geosciences, held by a university geology lecturer and five undergraduates at the beginning and end of the academic year. Participants were provided with six names of science subjects (e.g., physics, geography, geology) as elements and fifteen constructs were elicited by Kelly's triadic method separately at the beginning and the end of the year. Construct labels were treated as qualitative data and were classified into five categories by the author: (i) nature of science (e.g. *objective/subjective*); (ii) aspects of investigation (e.g. *use of maps or charts/no such use*); (iii) application of science and its professional aspects (e.g. *more employment opportunities/less*; *modifies environment/preserves environment*); (iv) affective aspects (e.g. *like/dislike*; *difficult/easy*); and (v) characteristics of the courses (*with lab/without lab*). Bezzi (1999) found constructs to be predominantly of the first two kinds, that is, dealing with the scientific essence of the disciplines. He used principal component representations of both elements and constructs (as in Figure 29.6) to identify constructs associated with the elements 'geology' and 'geography' both before and after the course of instruction for each of the five students and the lecturer to show how their perceptions of these two subjects had shifted over the year. Bezzi (1999) was able to conclude that simply studying the content of science did not lead to a greater understanding of the role of science in society or how to make good public decisions about scientific issues confronting society.

Lui and Lee (2005) showed how repertory grids could be incorporated into a learning programme in an examination of conceptual understanding in computer-mediated peer discourse. Twelve graduate students rated each of six database design methodologies on eleven supplied database design concepts. The course instructor also completed this task. Use of supplied elements and constructs enabled the researchers to feed back information to students about conflicts between their perspectives and for the students to take these differences into account during a week of online unstructured discourse in which students were required to reach a consensus perspective. Following this, students again completed the grids, and it was found that student concepts became significantly more aligned with the instructor's perspective. Lui and Lee (2005) conclude by suggesting that the methodology could be extended to using student-derived constructs, although this would require more complex procedures.

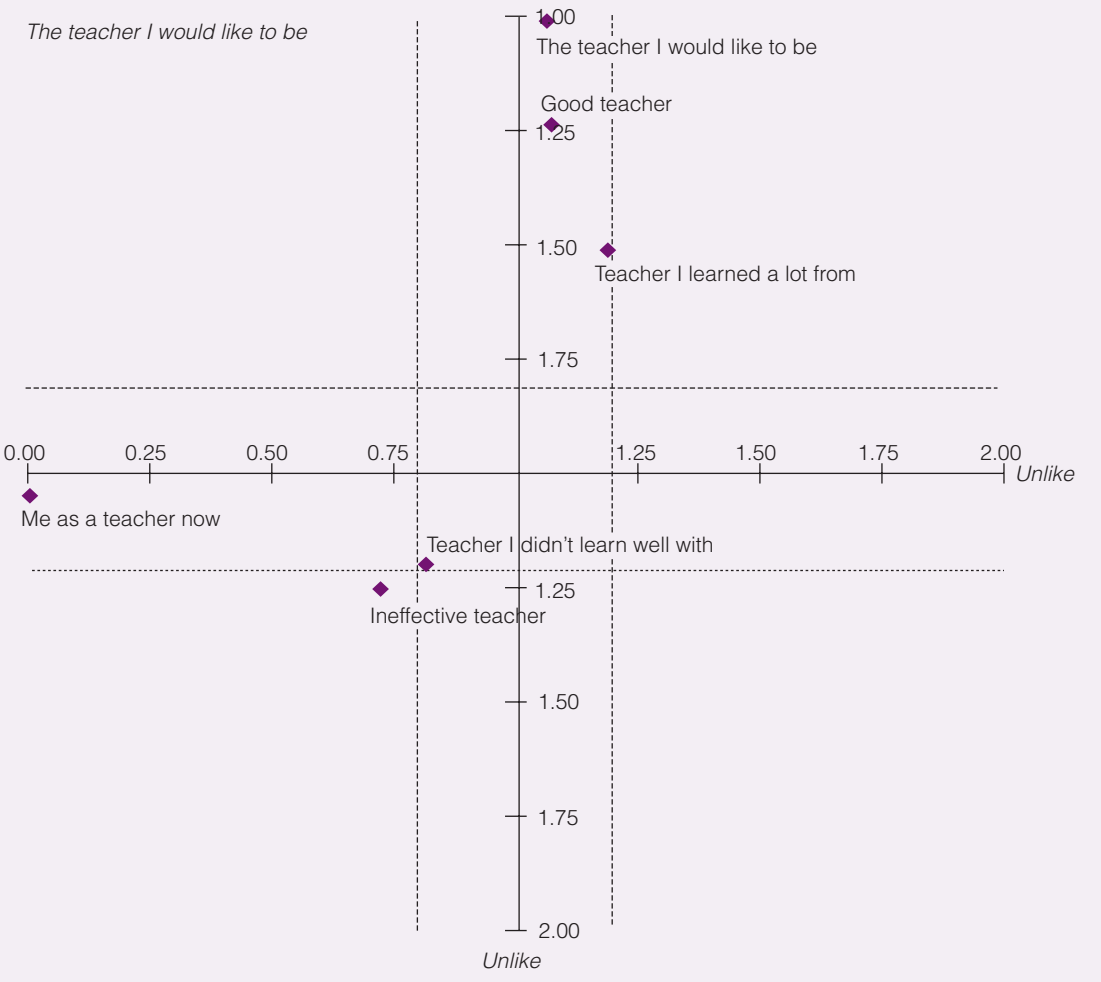


FIGURE 29.5 Self-identity plot

Yeung and Watkins (2000) employed the repertory grid technique to investigate how student teachers in Hong Kong developed a personal sense of teaching efficacy. A pilot study was used to generate elements such as 'self-efficacy', 'teaching practice', 'teaching practice supervisors', 'pupils' and 'lessons'. Constructs were individually elicited from twenty-seven students using the triadic procedure with cards. Yeung and Watkins (2000) matched constructs between student teachers to identify core constructs and create networks of similarity among student teachers using the Repgrid software of Shaw and Gaines (an earlier version of their current package, discussed above).

They found that third-year students' perceptions were more homogeneous than those of first-year students. Teaching efficacy was defined in terms of the dimensions of concern for instructional participation and learning needs of pupils, communication and relationships with pupils, academic knowledge and teaching skills, lesson preparation, management of class discipline, teaching success, teaching commitment and a sense of self-confidence. Experiences of teaching practice, pupils and teaching practice supervisors (Electives) were the major sources for the development of a sense of teaching efficacy.

Suto and Nádas (2009) used repertory grids to investigate why some GCSE examination questions in mathematics and physics were harder to mark accurately than others. Two highly experienced principal examiners generated constructs of question features for triads of ques-

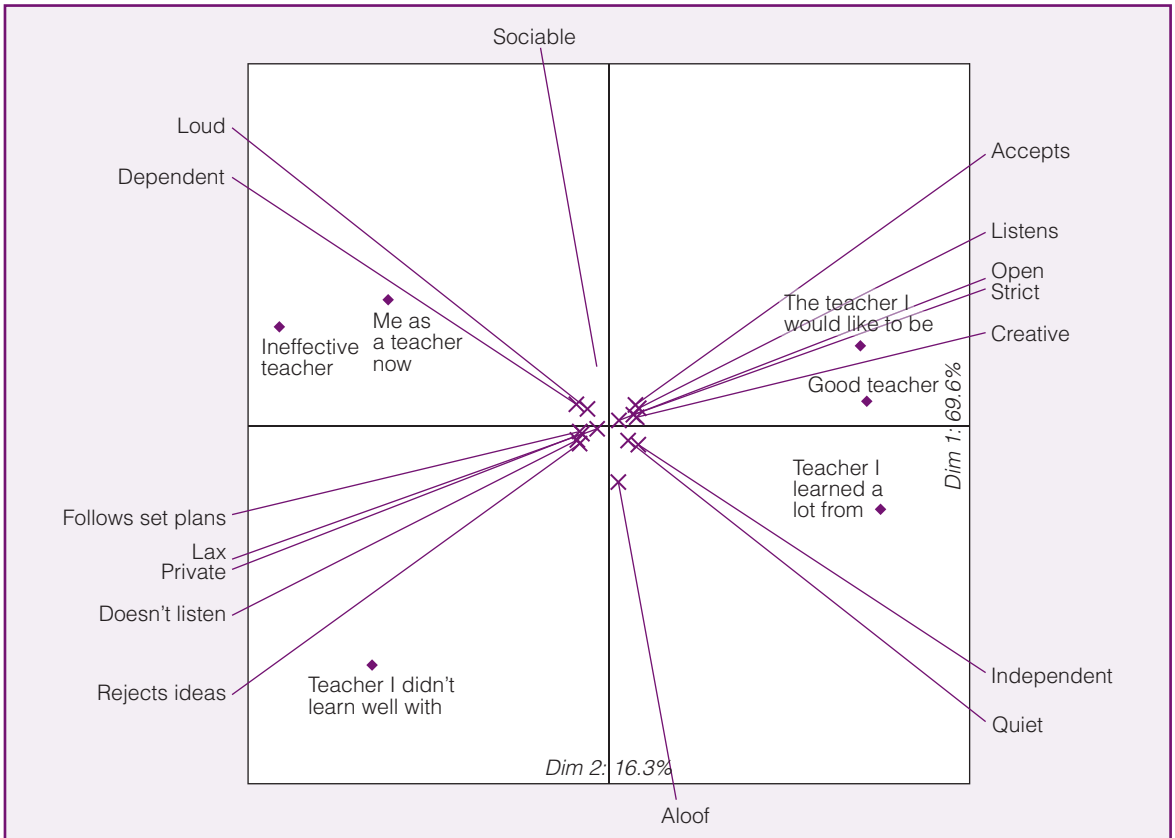


FIGURE 29.6 Spatial representation of elements and constructs

tions (elements) and rated each question (i.e. the elements) on these constructs. The study examined the constructs generated in detail and related this to marking accuracy. In a similar vein, Johnson (2008) used the repertory grid procedure with assessors of vocationally related portfolios to elicit constructs of differentiation among portfolios. Six assessors generated 131 constructs over six assessment objectives. There was general agreement between the assessors about the qualities of the commonly identified constructs, but Johnson (2008) did identify some potentially problematic linguistic issues, usually between the notions of quality and quantity.

Crudge and Johnson (2007) used the repertory grid to elicit a set of constructs relevant to web search engines (such as Google) from ten information science undergraduates. They then used laddering to determine the reasons for a construct's importance within the user's mental model. Using standard qualitative techniques they identified three hierarchical strata that conveyed the interrelations between basic system description, evaluative description and the key evaluations of ease, efficiency, effort and effectiveness. Two additional layers related to the perceived process and the experience of emotion are also discussed. They concluded that their model of key evaluations with the conjunctions of procedural elements provided a framework for further research to evaluate search engines from the user perspective.

In another use of laddering, Voss *et al*. (2007) used two laddering techniques (personal interviews and laddering questionnaires) to identify desired qualities of lecturers in a sample of eighty-two business management students. They found that personal interviewing led to more complex ladder structures with more components. Among the substantive findings they found that students' academic interests motivated them less than the vocational aspects of their studies.

Madill and Latchford (2005) explored identity change in four medical students over their first year of medical training, particularly in relation to their experience of human dissection. Each participant completed two repertory grids (one oriented towards their identity

construction, the other drawing out their experience of human dissection) at two time points, early in term one and towards the end of term three. The identity constructs elicited involved three common themes: dedication, competence and responsibility, as well as negative reactions, such as feeling driven and stressed. Three major themes were apparent in their experience of human dissection: involvement, emotional coping and ability. Complex patterns of relationships between the grids and between occasions led the authors to see a development of a vulnerable sense of professionalism alongside a frustration at potentially losing out on wider aspects of personal development.

29.6 Competing demands in the use of the repertory grid technique in research

The major overarching issue in repertory in research is the tension between individuality and commonality. A grid which is elicited wholly from the respondent is the most valid representation of that person's construing. Research, however, demands replication across subjects, so that for some purposes there needs to be commonality across respondents. Where the researcher's interests are qualitative, then individual grids form a useful way of collecting qualitative data. Some studies in the previous section provide examples of this feature.

When the structure of the grid data is of interest, the quantitative aspects of the grid become important. When these are specific to an individual because of the individualized specification of elements and elicitation of constructs, the quantitative component of an individual's grid cannot be related to that of others. In such situations the only traditional way these individualized grids could be compared is through the use of grid summary measures, such as 'cognitive complexity' or self–ideal discrepancy across constructs. When grids have some aspects in common, for example, role-specified elements or supplied constructs, it is possible to analyse the common aspects of such grids (see Fransella *et al.*, 2004, pp. 98–101). A recent development for data analysis in the individualized construct situation discussed above makes it now possible to carry out analysis at the level of construct. This can take account of within-grid variation at the simplest level, but where constructs can be commonly (across grids) qualitatively categorized, the categories can then be used as a factor in the analysis. Heckmann and Bell (2015) have recently shown how this can be carried out with free web-based software.

There are also some issues with aspects of grid elicitation, particularly relating to the issue of bipolarity in the grid. When only one pole of the construct is used, unwarranted inferences about constructs' polar opposites may be made. Yorke's (1978) illustration of the possibility of the researcher obtaining 'bent' constructs might suggest the usefulness of the opposite method (Epting *et al.*, 1971) in ensuring the bipolarity of elicited constructs. There is also the previously mentioned uncertainty about the meaning attached to mid-point ratings. Value-laden element role titles (such as 'A teacher I disliked') can affect the structure of the grid (see Haritos *et al.*), as can the orientation of construct poles. A number of practical problems commonly experienced in rating grids are identified by Yorke (1978):

- variable perception of elements of low personal relevance;
- varying the context in which the elements are perceived during the administration of the grid;
- halo effect intruding into the ratings where the subject sees the grid matrix building up;
- accidental reversal of the rating scale (mentally switching from 5 = high to 1 = high, perhaps because '5 points' and 'first' are both ways of describing high quality). This can happen both within and between constructs, and is particularly likely where a negative or implicitly negative property is ascribed to the pair during triadic elicitation;
- failure to follow the rules of the rating procedure. For example, where the pair has had to be rated at the high end of a five-point scale, triads have been found in a single grid rated as 5, 4, 4; 1, 1, 2; 1, 2, 4, which must call into question the constructs and their relationship with the elements.

Laddering also presents problems, both in process (Butt, 1995) and in the hierarchical implications that follow (van Rekom and Wierenga, 2007).

Another problem is the continuing tension between theory and method in the repertory grid. Major works devoted to the repertory grid technique (Bannister and Mair, 1968; Fransella and Bannister, 1977; Jankowicz, 2003; and Fransella *et al.*, 2004) have all been written from a personal construct theory perspective and emphasize the importance and relevance of Kelly's theory to the usage of the technique. Yet most research with the repertory grid is carried out with the grid being used in a purely methodological and a-theoretical fashion that is content with a passing reference to Kelly as the originator of the grid. As this chapter demonstrates, the theory can be used to understand what is happening in the grid, but of course it is not essential to its use. The real drawback to the personal construct

theory background of writers in this area is often with the jargon employed, particularly for indices or measures developed to summarize structures in the grid. For example, the average correlation used to summarize relationships between constructs was originally termed 'intensity' by Bannister (1970) and is often referred to by that name.

29.7 Resources

It might be thought that the repertory grid is unchanging. However, there has been continuing research to inform the technique itself (some details of such developments have been mentioned in this chapter) as well as new measures that can be derived from grid data (for example, Bell (2004b) devised a new index of inconsistency in ratings in the grid) that could have applications in educational research. Most, if not all, of these developments occur through the work of researchers identified with personal construct psychology. Such researchers tend to work alone or in small groups across the globe and use the web as a means of communication. There is a very general and comprehensive site maintained by the George Kelly Society with many resources and links to groups, publications and computer programs located at www.kellysociety.org/resources.html.

The major computer programs for the analysis of grid data through SPSS can be found on the Kelly Society website (http://kellysociety.org).

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Role-play and research

CHAPTER 30

Carmel O'Sullivan

This chapter introduces role-playing as a research technique, and discusses issues such as:

- role-play pedagogy
- what is role-play?
- issues to be aware of when using role-play
- role-play as a research method
- how does it work?
- strategies for successful role-play
- a note on simulations
- examples of research using role-play

Role-play is very valuable for researching questions such as 'what if', 'how', 'why', 'if–then', 'what are the effects of', 'what are responses to', 'what are the challenges in', 'what happens if', 'what are the key issues in'. They can examine individual and group behaviour in controlled and less controlled environments, in safe environments and in particular situations. They are useful in studying complex situations, interactions and evolving decisions, and in which issues capable of different interpretations, negotiations or potential conflicts are included.

30.1 Introduction

Erving Goffman (1969, p. 78) famously claimed that '[l]ife itself is a dramatically enacted thing', and although he recognized that 'all the world' is not a stage, he argued that 'the crucial ways in which it isn't are not easy to specify'. The perception of people playing some role or other in their everyday lives is a common theme in literature. The Irish playwright Oscar Wilde similarly advanced the argument that life and drama are closely connected:

> Actors are so fortunate. They can choose whether they will appear in tragedy or in comedy, whether they will suffer or make merry, laugh or shed tears. But in real life it is different. Most men and women are forced to perform parts for which they have no qualifications. … The world is a stage, but the play is badly cast.
>
> (Oscar Wilde, *Lord Arthur Savile's Crime*, 1891)

It is undeniable that drama, in some form or other, touches most people's lives, typically in society today through watching television soap operas, going to the movies or attending the theatre. However, drama has been recognized for many years as a useful training method in the fields of business, health sciences, industrial psychology, law, sociology, political science and education. Harriet Finlay-Johnson (1912) and Henry Caldwell Cook (1917) were using dramatic play in schools as a teaching and learning method in England at the turn of the twentieth century, while Jacob Levy Moreno (1908) was similarly exploring its use with children in Vienna, before later developing its application in therapeutic procedures known as psychodrama (Moreno, 1939).

Owing to the immediacy of its impact, drama can affect people in powerful ways, and historically it has been viewed unfavourably by ruling hegemonies wishing to suppress its ability to move audiences and mould attitudes (Banham, 1995). As is the case with all art which has the power to 'move people', the educational use of drama has the potential to connect with people both emotionally and cognitively, resulting in what we might call 'felt understanding'; a type of knowing which results in people taking a personal interest in issues and wanting to effect change. This potentially subversive power has been recognized throughout the ages by different ruling elites and oppressive regimes, who sought to diminish or eradicate its power through banning or severely censoring it.

Fearing the connection between political radicalism and social unrest, and the powerful role of theatricality in everyday life (Nicholson, 2015), censorship of the theatre and dramatic performativity in people's daily lives remained active throughout most of the twentieth century. However, recent archival and other evidence reveals that bureaucratic regulation, prohibition and state control in right-wing dictatorships, post-colonial regimes, communist systems and western democracies continue to censor drama and theatre performances in the twenty-first century (O'Leary *et al.*, 2015).

In spite of this, drama's potential as a valuable educational method has endured, and with the publication

of Maier *et al*.'s training manual for role-playing in 1957, an ever-increasing number of researchers have been motivated to use role-play in their research design. Described as 'a pioneering attempt to portray industrial conflicts in role-playing format' (Mee, 1957, p. 135), the authors (Maier *et al*., 1957, p. xi) recognize that understanding the 'principles of human behavior has little value unless it is supplemented with skill practice'. Maier *et al*. (1957) describe one of the benefits of using role-play as being able to demonstrate the gap between thinking and doing. The innovative approach was well received at the time by researchers in a number of applied areas, including sociology, education, management science and industrial relations (see Berg, 1957; Borgatta, 1957; Mee, 1957; Argyris, 1958), and researchers recognized in this fusion of case study method with role-playing techniques a rich potential to analyse aspects of social behaviour and social interaction implicit in ordinary living.

In the succeeding sixty years, role-play has been widely used in education and training, and increasingly in the field of corporate training, where it is used to explore such issues as change management, negotiation skills, communication skills, leadership skills, entrepreneurial attitudes, team building, presentation skills, management training, public speaking, assertiveness training, performance management, customer service, interview skills, stress management, appraisals training and media training.

Role-playing, gaming and computer simulation are three related strands of activity in this wider field, but it is beyond the scope of this chapter to explore the last two categories in detail (but see Chapter 23 of the present volume), and the focus here is on the use of role-play as a technique in educational research. Similarly, the area of online role-play and game-based learning is currently burgeoning, with many people beginning to research the areas of socialization and social identity theory as key constructs underpinning learning in massively multiplayer online role-playing games (MMORPG) (see Liang, 2012; Moon *et al.*, 2013). These experiences, while usually recreational, can also offer valuable social and educational outcomes. Moser and Fang (2015) refer to the value of narrative structure and decision-making points as improving participants' experience of online role-playing games.

Prosocial behaviour and skills, absorption and empathy are also positively identified with online role-playing games (Rivers *et al.*, 2016). Research is indicating that role-playing can be effective in online language learning and teaching (Ashraf *et al*., 2014; Zhang *et al.*, 2016). The developing research area of how to respond to 'unexpected events' also currently features as a key characteristic of digital game-based learning. The creation of immersive virtual worlds which facilitate participants' realistic interactions and enhance their experiential learning of human-related issues, such as teamwork, communication and collaboration, is attracting much attention in the literature (Maratou *et al.*, 2016; Sampson *et al.*, 2014) (see Chapter 23 of the present volume).

30.2 Role-play pedagogy

Education systems worldwide are constantly reinventing themselves in the twenty-first century in an effort to keep pace with the exponential growth in digital learning and media. Technology is dramatically changing the nature of learning, training and education, and voraciously demanding rapid sectorial and societal responses. Terms such as project-based learning, STEM/STEAM, blended learning, flipped classrooms, inquiry-based instruction, mLearning, T-learning, microlearning, social learning, gamification, Tin Can and personalized learning, all feature as the latest buzzwords in the field of learning and development.

Experiential learning is quickly becoming an umbrella term for many of these ideas where the emphasis is on active and creative forms of learning that provide learners with experiences through which they can acquire, experience and develop new knowledge and skills (on- or offline). It is within this context that the idea of ownership and 'owning learning' has resurfaced in education as a response to technology-assisted learning and the changing profile of many learners globally.

There is a common theme running through much of the recent literature in education which reflects the idea of empowered, autonomous, active, participatory and creative teaching and learning practices for lifelong learners in the new millennium, and role-play is being called into service as a way of encouraging engaged, enjoyable and deep learning from early years to adult education. With the current emphasis on experiential learning, role-play is seen as motivational, helping to develop student confidence and facilitate graduate skills acquisition (Waters, 2016).

Recent studies across most disciplines are reporting greater success in inviting learners of all ages to engage in role-play, and it is increasingly featuring as a core teaching, learning and assessment strategy across all levels of the education and work systems. Terms such as role-play pedagogy are regularly cited (Shapiro and Leopold, 2012; Gordon and Thomas, 2016), and it is explicitly associated with pedagogic aims placing

students in real-world contexts and encouraging them to explore the complexities of decision making, reflecting on issues and the views of others (Smith, 2016). Reflection features consistently in the literature on role-play, and Wadensjö (2014) found it effective as a method to encourage examiners to reflect on role-play as an assessment instrument in examination sessions (see also Johansson *et al.*, 2012).

In keeping with the emphasis on what are loosely termed twenty-first-century skills, many of which are associated with social and emotional learning, role-play is increasingly being used in multiple formats:

- to 'stand in the shoes of someone else' (role-switch: to learn from the inside out);
- acting (practising a skill); and
- 'almost real life' (experiencing something as close to the real-life example as possible) (Rao and Stupans, 2012).

This aligns closely with the current emphasis on social skills, metacognition and lifelong learning, particularly with 'millennials' and the newer 'generation tap', who are born in a digital age where vast amounts of information are available at a touch, but where many learners need to learn how to use, critically interpret and apply it.

Ideas such as adding variety, interest and involvement in learning, increasing students' sensitivity to feelings and the attitudes of others, enhancing the authenticity of experiences, giving students insight into the dynamics of interpersonal interactions in relationships which cannot be gained from other methods, were commonly associated with role-play in the 1950s and 1960s (Buxton, 1956; Shipman, 1964) and are making a strong comeback as part of the twenty-first-century suite of skills, which arguably is connected to role-play.

Although role-playing is increasingly featuring in the literature in many subject areas as a creative and participatory teaching and learning strategy, it has had mixed success to date as a research method, much of it owing to a persistent lack of clarity about what it is and how to use it, as revealed in an early report from the *British Medical Journal* in Box 30.1.

We begin by taking a closer look at drama and role-play in order to increase our understanding of what it is and how it works.

30.3 What is role-play?

Deriving its theoretical basis from the field of psychodrama, role-play is a 'spontaneous, dramatic, creative teaching strategy in which individuals overtly and consciously assume the roles of others' (Sellers, 2002, p. 498). Sellers argues that it involves 'multi-level communication' (p. 498), and as a powerful teaching strategy is capable of influencing participants' attitudes and emotions, whilst simultaneously promoting higher-order cognitive skills. This definition supports the claim that role-play is an effective strategy for learning because it encourages participants to think about the person whose role is being assumed, is connected to real-life situations and promotes active, personal involvement in learning (Billings and Halstead, 2009). Errington (1997, p. 3) defines role-play as 'a planned learning activity designed to achieve specific educational purposes'. He suggests that it is based on three major aspects of the experiences that most people have of roles in everyday life:

- role-taking (the roles we hold in accordance with social expectations and in social circumstances, i.e. how police officers should act) (Goffman, 1976);
- role-making (the ability to create, switch and modify roles as required) (Roberts, 1991);
- role-negotiation (negotiation and social interaction with other role holders) (Hare, 1985).

For educational researchers, these categories offer a wealth of possibilities for accessing and exploring people's behaviour and responses to situations and stimuli in a diverse range of contexts and settings. For example, a researcher investigating a new coaching and mentoring training approach for senior managers in schools may involve participants in varying aspects of role-taking, role-making and role-negotiation as part of his overall research design to gather relevant data.

Role-play consists of three major stages: briefing, acting and debriefing. The first stage focuses on introducing the participants to the activity by clarifying the learning objectives and 'setting the scene'. In the second stage, the educator must encourage the participants to 'act out' the role in a spontaneous, accurate and realistic manner. Debriefing, the final stage, allows participants to discuss, analyse and evaluate the role-play and insights gained (Billings and Halstead, 2009).

Working in drama involves stepping into an imagined world, a fictional reality, and in order to make this imaginary world more meaningful and purposeful in an educational research context, it must have aspects of the real world in it. Thus, human relationships are a central component of role-play situations, and exist in the form of:

1. relationships between people;
2. the relationship between people and ideas; and

BOX 30.1 A ROLE-PLAYING EXPERIENCE

It's Wednesday morning again, and time for our clinic sisters' teaching session. Anybody who thinks this gives me a relaxing two hour break from the rigours of outpatients is sorely deluded. Seeing a hundred patients seems quite a soft option compared with facing our four most senior sisters, exercise books open and pens poised to take down my every word.

Thinking up suitable topics is not easy. But harder is the actual task of teaching. British medical training provides ample case studies in how not to teach, and I've wasted many hours trying to find a comfortable sleeping position while a well intentioned lecturer starts on yet another new piece of chalk. Surely I can do better than that?

So I've put away the blackboard. And I've got the chairs rearranged in a circle. Dividing up for group work is tricky when there's only five of us all together. 'Brainstorming' with the flip chart is a bit of a non-starter with a group that's as talkative as a bunch of Trappist monks. But for today's session, on AIDS counselling, there is only one possible option. I must introduce them to the joys of role-play.

After a prolonged discussion about counselling in general, we kick off with a simple scenario. I am heading the bill with a stirring performance as Sipho, a young Zulu man who will need to be told that he is HIV positive. I have been rehearsing my lines for sometime, and I am all ready to bring the audience to its feet with my impassioned soliloquy. Sister Gumede has bravely volunteered to star in the lead role as, well, a clinic sister. After all, it is the first time they have ever done role-play, and I don't want to put them off.

So I reel off my performance, standing up, gesticulating, groaning, and clutching my head as I hear the bad news. Only the glycerine tears are missing. But the audience is not moved. They watch with bemused perplexity, and take copious notes. Sister takes her cue, and improvises her lines deadpan: 'Sipho, your HIV test is positive. We don't have any cure for AIDS so you are going to die.'

Now I have real cause to groan and clutch my head. What sort of a way is that to counsel someone who is HIV positive? Where are the open ended questions, the active listening, the non-verbal communication? My Balint colleagues in Lisson Grove would throw up their hands in horror.

But this is Africa, not north London. Maybe sister's performance is the one that should win the Oscar. After all it is exactly the way most of the sisters talk to patients, and probably the way that patients expect to be spoken to. And whoever saw a Zulu man behaving anything like my performance? Maybe this role-play business is not so simple. Maybe teaching is not so simple. So next time, sisters, bring your pillows; I am going to write notes on the blackboard.

Duncan Curr, Medical Officer, Mosvold Hospital, Ingwavuma, South Africa

Source: Curr (1994, p. 725)

3 the relationship between people and the environment (O'Toole and Haseman, 1992, p. 3).

These categories provide a useful framework for researchers interested in using role-play to identify who and what their research should focus on. Thus a researcher who wishes to explore whether teenagers empathize with bullied peers may use role-play to determine the extent to which young people:

1 discuss the issue among themselves, and if/how they would approach the subject with a peer in school who they know is being bullied;
2 engage with training sessions and resources they have received in school on the issue of bullying, and if/how they put these into practice;
3 demonstrate an awareness of their role in the creation of a culture that does not accept or tolerate bullying in their school environment and wider social community.

Taking on roles allows participants to set up and explore the different dynamics of the relationships cited above, but differs from traditional understandings of theatre in that the role-taker is not required to demonstrate elaborate acting skills, rather, simply to represent a point of view. Ideally, the role should be portrayed honestly and without elaborate costumes or props, where participants place themselves 'as if' they are that person, temporarily identifying with and exploring a set of attitudes and values, which may not identify closely with their own. It is therefore important that

participants respect the role being played, as it represents another person's perspective or point of view (O'Toole and Haseman, 1992, p. 3). Lowenstein (2016) describes role-play as a versatile technique which focuses on the actions of the characters, their attitude to the situation, and not on their acting ability. The role-play must be lived at life-rate (i.e. in that moment), and aim to create a living picture of life, which provides a learning opportunity for the participant as much as for any onlookers, including the researcher.

The more realistic the role-play, the better the outcomes, according to Dennison (2011), who found in her study that using theatre studies students as patients was not as realistic as using a more traditional approach to role-playing where the social work students played the role of the patients themselves.

Role-play is improvisational in nature and increasingly unscripted, although role-cards may be supplied to provide sufficient background information to participants to enable them to comfortably 'step into the shoes' of another, and feel what it might be like to be that person in that situation for a little while. Wagner (1998, p. 60) defines improvisational drama as taking on 'a role in a particular moment in time and creating with others a plausible world'. She argues that when working in role, as in all learning contexts, participants make meaning by connecting their prior experiences to the challenge of the moment (see Bradshaw and Hultquist, 2017). However, using a non-parametric test, Osborn and Costas (2013) found no statistically significant difference in the development of counselling skills between students improvising their own role-plays and using scripted role-plays.

30.4 Why use role-play in research?

Cabral (1987, p. 470) describes role-playing as a valuable technique that 'has been broadly adapted for use in academic research and applied settings'. Before presenting the arguments related to the particular use of role-play in research contexts, it is worth taking a brief look at the general educational claims made in its name. In their aptly titled book *So You Want to Use Role-Play?*, Bolton and Heathcote (1999) provide six major categories which summarize the use and value of role-play in education.

1. Behaviour modification. It is a concrete form of learning, and particularly suitable for giving participants practice in behavioural procedures (for example, training reception staff or police officers to handle particular situations according to an established procedure).
2. Acquiring information.
3. Using information.
4. Training in seeking information.
5. Attention to detail (role work can generate in participants a disposition to attend to detail, an alertness to particulars).
6. Fostering a change in values, perceptions or attitudes.

(Adapted from Bolton and Heathcote, 1999, pp. 178–85)

The arts, and role-play in this particular discussion, work by revealing truths about people and the world they live in, and they do this through the creation of a fictional situation, a make-believe world, but one that is closely connected to reality. This is the difference between fiction and fantasy, and in most research situations where role-play is used, the emphasis is on working in and through fiction: uncovering and exploring truths about reality, and about how we respond individually to such situations, as we each construct our own understanding of experiences. Bolton and Heathcote (1999, p. ix) are concerned with broadening the traditionally perceived use of role-play, away from a strictly behavioural emphasis to the communication of meaning. Thus, an educational researcher interested in investigating social skills education with children with an autistic spectrum disorder may use role-play to create a baseline assessment of participants' 'theory of mind' (i.e. their ability to perceive and understand the thoughts of others) through placing the children in role as detectives, observing a crime as it unfolds and trying to predict what the characters are thinking at that time (see O'Sullivan, 2015).

There are many uses and types of role-play, but the single criterion underpinning all role-play activity, according to Bolton and Heathcote (1999, p. 57), is that it demands participants to step into an 'as if' fiction: a fiction that has been 'conceived of by a tutor, teacher, or researcher in terms of *learning*'. It poses a unique challenge to participants, as it involves 'embracing knowledge', an act which is not just a matter of instruction or absorption, but is achieved by entering the fiction in such a way as to make the required knowledge one's own. Thus, role-players are not just receiving or acquiring knowledge as in a typical instructional context; they are making it, practising it and embodying it: they know what they know (Bolton and Heathcote, 1999, pp. 57–8). This highlights the importance of accurately setting up and structuring the role-play to record these truths and attitudes, and thereby increase the reliability and validity of the data retrieved.

Roslyn Arnold (1998, p. 111) claims that the arts, and in particular drama, uniquely explore the dynamics between affect and cognition, two significant aspects of human existence. The use of role-play in research contexts is a rich source of insight into the role and function of such dynamics. Thus, when participating in role, we articulate both physically and verbally, and this active engagement promotes 'emotional, cognitive, social and ego development. We are, metaphorically speaking, sitting on a research gold mine' (Arnold, 1998, p. 111).

One of the main reasons for considering the use of role-play in research is because of its ability to help participants consider ideas from different perspectives, to think of possibilities. Role-play is concerned with representing and exploring different people's points of view, and different points of view forge different types of knowledge. It places participants at the centre of the learning experience, and allows them to build their own bridges of understanding. As a result of this informed consideration, they are better able to resolve problems and issues. For example, role-play as a research method has been successfully used with young people in a designated disadvantaged school to elicit the extent to which they use an elaborate or restricted linguistic code (public versus private use of language), and whether, through the use of role-playing, they can explore and develop different linguistic registers and codes as appropriate to a range of communication contexts (Heeran-Flynn, 2010).

The following list identifies the range of possibilities in which role-play can be employed as an effective research method. Specifically, role-play can allow participants to:

- experience how people behave in particular circumstances by exploring a variety of social situations and social interactions;
- explore a range of human feelings and responses to situations;
- explore choices and moral dilemmas;
- make decisions which are tested out in the role-play and later reflected on;
- develop a sense of responsibility and confidence as decision makers and problem solvers;
- improve the social health of their group and foster improved relationships with peers or colleagues;
- interact with peers and learn to compromise in order to sustain and develop activities;
- extend, enrich and prompt the use of authentic language use in simulated real-life contexts where language use arises out of a genuine need to communicate;
- explore the skills and processes involved in conflict, negotiation and resolution of difficulties and problems in their environment;
- develop personal creativity;
- develop agency and an increased awareness of self;
- improve visual and spatial skills through responding to a range of stimuli and situations.

Role-play situations as described above can be observed by the researcher, and/or digitally recorded, and replayed to participants to elicit their responses and perspectives according to predetermined or emerging research themes and issues, thereby assisting in the triangulation and interpretation of data. Such an approach was used when investigating social skills education with children and young people with Asperger's syndrome. Role-play was used as a core research method in a longitudinal study to initially create a baseline measure of participants' literal and metaphorical language competencies, and then to assess the extent to which improvements in participants' social skills were revealed and practised during subsequent role-play episodes (see O'Sullivan *et al*., 2009).

Role-play operates in a 'no-penalty zone', where people are freer to explore and try out a range of solutions to problems and issues, without having to worry about the outcome. Such an approach was adopted in a recent study designed to explore the impact of role-plays on primary school children's awareness and knowledge of bullying in an all-boys school (Donohoe and O'Sullivan, 2015). Using the *No Blame Approach* (Maines and Robinson, 1997), within the framework of a role-play-based *Bullying Prevention Pack* (BPP), a standardized survey instrument (Olweus Bully/Victim Questionnaire – Revised, OBVQ-R) revealed a 53 per cent reduction in reports of victimization at the research school.

Drama functions as a way of making the world simpler and more understandable. It can be a kind of 'playing at' or practice of living in real-life situations. It enables participants to put into practice skills they have learnt in the fictional context of the drama world. It is a tool that can affect participants' fundamental reactions to everyday situations. Augusto Boal (1979, 2002) refers to this type of dramatic activity as a 'rehearsal for reality'. In professional disciplines, such as health care, education, engineering and social care practice, both pre- and in-service education models rely on problem-based and enquiry-based approaches to teaching, learning and research. Role-play offers enormous potential in these areas to enhance case study method and facilitate research on models of best practice.

The notion of learning through play tends to be associated with early years education, and opportunities for imaginative and dramatic play decline as a child progresses through the school system and into adult life. An unfounded belief that academic content standards cannot be met through creative and imaginative activities still persists, and has caused playful methods of learning to virtually disappear from classrooms (O'Sullivan, 2016a). The following is a brief summary from the literature reflecting the use of role-play in research on a continuum of lifelong education.

- Role-play facilitates participatory research in case study method (Carte and Torres, 2014).
- Role-play provides rich opportunities to assess communication skills and language use in simulated training sessions (Stokoe, 2013, 2014).
- Role-play is an effective strategy in exploring ethics (Strohmetz and Skleder, 1992; Doron, 2007; Kraus, 2008; Roos, 2011).
- Within a socio-linguistic constructivist approach, role-play can help learners access and communicate abstract ideas in science (McEwen *et al*., 2014; Braund *et al.*, 2015).
- The integrated nature of role-play allows for individual differences in development (Frost *et al*., 2008).
- Role-playing is a popular teaching method with students, and encourages participation, engagement and active learning (Stevens, 2015; Waters, 2016).
- Role-play activities help participants create more informed opinions and stimulate critical thinking and argumentation skills (Agell *et al*., 2015).
- Effective classroom management strategies can be explored through role-play in teacher education programmes (Niemeyer *et al.*, 2014).
- Online role-play is effective in language instruction and the development of collaborative argument (Zhang *et al*., 2016).
- Complex human–environment relations and environmental governance issues can be explored effectively through role-play (Schnurr *et al*., 2014, 2015; Agell *et al*., 2015).

30.5 Issues to be aware of when using role-play

Much of the early history relating to the use of role-play in research settings was mired in controversy and notoriety, mainly relating to issues around deception in experimental social psychology (see Milgram's obedience to authority experiments, 1974; Mixon's role-playing replications of the Milgram experiment, 1974), and to overt/covert forms of research. Bolton (1996, p. 187) discusses the case of James Patrick (a pseudonym), a young teacher at an approved school who in the late 1950s obtained entry into a Glaswegian gang for four months, in order to record and analyse how a city gang functions (see Patrick, 1973). He made friends with a pupil in his school called Tim, and through this acquaintance, joined his pupil's gang. In deciding to open his teacher's eyes to gang life, Tim understood the risks more clearly than Patrick did. Tim was extremely well-behaved when in school, but at weekends he participated fully in the violent incidents that regularly erupted at a moment's notice (Douglas Home, 2007). After having been placed in several uncompromising situations, Patrick left Glasgow quickly when the violence became too severe and he felt threatened by it. As he was so afraid of the gang members, he didn't publish his research until many years later.

Bolton (1996) describes this act of infiltration and deception as a blatantly unethical form of inquiry, although a researcher may well have to ask themselves whether the same information could have been gained by overt means. In contrast to this covert approach, William Foote Whyte (1993) conducted a similar research exercise in a poor Italian district in Boston from 1937 onwards called 'Street Corner Society' (see also Chapter 35 of the present volume). He was interested in the activities of the adolescent boys who hung around street corners and got involved in gang activities. However, his research approach was not covert, and he began almost as an observer, having informed the group that he was writing a book about their activities (Whyte, 1993). Perhaps one of the most controversial examples of a study involving the use of role-play is the well-known Stanford Prison Experiment carried out by Philip Zimbardo in 1971 (2000, 2007a, 2007b, 2008; see also Haney and Zimbardo, 1998), a brief overview of which is given in Box 30.2.

The Stanford Prison study raises uncomfortable questions for researchers. On the one hand, it violated the principle of 'do no harm' (see Chapter 7 of the present volume), exposing the 'prisoners' to cruel behaviour, suffering and sadism, to psychological torment and distress, to the removal of freedom and self-control, to degradation, to the loss of identity in the prisoners, to physical abuse, embarrassment (e.g. strip searches), harassment, public humiliation, to feelings of helplessness and despair, to emotional breakdown and to some longer-term trauma. As one participant said, years later: 'It is still a prison to me. … It harms me.'

On the other hand, the participants were volunteers who had undertaken diagnostic psychological screening

BOX 30.2 THE STANFORD PRISON EXPERIMENT

The study was conducted in the summer of 1971 in a mock prison constructed in the basement of the psychology building at Stanford University. The subjects were selected from a pool of seventy-five respondents to a newspaper advertisement asking for paid volunteers to participate in a psychological study of prison life. On a random basis, half of the subjects were assigned to the role of guard and half to the role of prisoner. Prior to the experiment subjects were asked to sign a form, agreeing to play either the prisoner or the guard role for a maximum of two weeks. Those assigned to the prisoner role should expect to be under surveillance, to be harassed, but not to be physically abused. In return, subjects would be adequately fed, clothed and housed and would receive fifteen dollars per day for the duration of the experiment. The outcome of the study was quite dramatic. In less than two days after the initiation of the experiment, violence and rebellion broke out. The prisoners ripped off their clothing and their identification numbers and barricaded themselves inside the cells while shouting and cursing at the guards. The guards, in turn, began to harass, humiliate and intimidate the prisoners. They used sophisticated psychological techniques to break the solidarity among the inmates and to create a sense of distrust among them. In less than thirty-six hours one of the prisoners showed severe symptoms of emotional disturbance, uncontrollable crying and screaming, and was released. On the third day, a rumour developed about a mass escape plot. The guards increased their harassment, intimidation and brutality towards the prisoners. On the fourth day, two prisoners showed symptoms of severe emotional disturbance and were released. On the fifth day, the prisoners showed symptoms of individual and group disintegration. They had become mostly passive and docile, suffering from an acute loss of contact with reality. The guards, on the other hand, had kept up their harassment, some behaving sadistically. Because of the unexpectedly intense reactions generated by the mock prison experience, the experimenters terminated the study at the end of the sixth day.

Source: Adapted from Banuazizi and Movahedi (1975)

before being admitted to the experiment, and they were provided with long-term follow-up support for many years. Further, its findings had, and continue to have, important implications: 'ordinary' everyday people have the potential to become sadists and to become highly emotional; the 'power of the situation' can exert extreme effects on people's behaviour beyond what might have been imagined, and indeed overtakes the power of the individual; the pathology of prisons is exposed clearly; and control and domination have rapid and powerful effects on all participants (e.g. prisoners and guards). In other words, the benefits to society were immense, and indeed spawned research into shyness and shyness therapy. Decades later, the experiment's director, Zimbardo (2007a) indicated how the 'Lucifer Effect' (his term) continued to operate in contemporary situations (e.g. in the US military's treatment of prisoners). The role-play provided findings which might not have been possible otherwise. The dilemma for researchers using role-play here raises awkward questions: Is the benefit worth the cost? Does benefit to the many override the harm to a few? Does the end justify the means? Further, for researchers, the Stanford Prison Experiment demonstrates very clearly the power of role-play as a research technique.

In discussing the Stanford Prison Experiment, Bolton (1996, p. 188) argues that the disregard of ethical standards in this research results from 'the tacit permission that role-playing a power-position gives', and not from deception. Although the researchers anticipated the risk of physical abuse and changed the rules to reflect this, they failed to predict the pleasure that some guards might derive from employing psychological abuse, 'even when they could perceive … the genuine discomfort of their victims' (Bolton, 1996, p. 188).

Early enthusiasts of role-playing as a research methodology cite experiments such as the Stanford Prison Experiment to support their claim that where realism and spontaneity can be introduced into role-play, then such experimental conditions do, in fact, simulate both symbolically and phenomenologically the real-life analogues that they purport to represent. Such advocates of role-play would concur with the conclusions of Zimbardo and his research associates that the simulated prison developed into a psychologically compelling prison environment, and they, too, would infer that the dramatic differences in the behaviour of prisoners and guards arose out of their location in different positions within the institutional structure of the prison and the social psychological conditions that prevailed there, rather than from personality differences between the two groups of subjects (see Banuazizi and Movahedi, 1975; Stokoe, 2013).

Bolton (1996) expresses concern that the use of role-play in such circumstances can appear to give permission to participants to behave outside their normal moral constraints. Grumet (1998) acknowledges that taking on a role is complex and provocative, and when working in role, researchers should be aware of 'the power of a role to extend or constrict meaning and exploration' (p. 8). If we accept the Latin word for role as *dramatis personae*, there is a danger that the participant may hide behind the mask of a role, taking on 'actions and ideas that would be difficult to assume within his or her daily identity' (p. 8). Thus, playing a role in this context may result in behaviours and attitudes that extend imagination and expression beyond the individual's usual capacity, and 'the imaginative extension of ego into role' might have the effect of constraining rather than enlarging understanding (p. 9).

On the other hand, Grumet (1998, p. 9) argues, when a role is used in a naturalized scene, untrained participants may 'fail to fill it with the complex and multiple possibilities that a real life situation' would demand, choosing instead to adopt a more stereotypical action in the improvisation than might exist in real life. The researcher must therefore look for opportunities to break the often powerful grip of a scene and role, and encourage critical reflection on the choices that are being taken in the role-play. The aim is to shift, alter, interrupt and possibly distort the focus of the role-play (during it if necessary), to allow participants to explore and experience different aspects of the situation under consideration, thereby 'avoiding a reductive metonymy that would substitute the improvisational situation for the world', with its infinite colour and myriad possibilities (p. 9). This can often be achieved by following step 8 (the hidden objective) as described in Box 30.3.

Wagner (1998, p. 58) suggests that working in drama requires the same intelligence it takes to live one's life in the real world, in order to be able to cope with the many possibilities, choices, decisions, ambiguities, changes, etc. that face people on a daily basis. The challenge in drama is to engage with those issues without losing the capacity to analyse situations responsibly and carefully, choose between alternatives that are not always clear, 'act on those choices and live with consequences. In other words, to think before, during, and after one acts' (p. 58). For the researcher interested in exploring the intricacies and complexities of life, the key is to set up and organize the role-play event so that it accurately reflects, not mirrors, the situation under scrutiny (see the guidelines provided in Box 30.3). When working with medical students, Joyner and Young (2006) highlight the importance of defining the learning objectives and setting the ground rules from the outset if a role-play is to be effective.

In using role-play in research trials such as the Stanford Prison Experiment, where none of the preliminary documentation given to participants refers explicitly to the act of role-playing, or provides them with information or guidance on how to safely 'enter a role', 'step into the shoes of another person' and behave as if they are that person for the duration of the activity (see www.prisonexp.org for copies of the original documentation), there is an implicit assumption that the person in charge of organizing the game is taking on ultimate responsibility for what might happen. Bolton (1996, p. 188) suggests that this effectively provides temporary release to the participants to regard the experiment as 'only a game and, what's more, someone else has asked us to play it'. However, he makes an interesting observation when he claims that this seeming release from responsibility rarely extends to breaking the rules of a game, and he recommends that researchers interested in using role-play may need to pay particular attention to this by 'delineating participant goals and delimiting strategies' (p. 188). It may have been a very different prison experiment had the responsibility been altered during the role-play by telling the warders that they had been nominated for promotion within the prison service on the basis of their ability to combine authority and respect in their dealings with prisoners (p. 188). The flexibility of role-play as a research method can provide a researcher with valuable opportunities to shift variables and explore other angles/perspectives, all within the framework of the existing role-play, thus saving time and resources (with due attention being paid to ethical issues and constraints as relevant).

In High Fidelity Patient Simulation (HFPS), which uses interactive computerized mannequins, priority is given to safety when teaching about complex clinical situations that mimic reality (Jarzemsky *et al.*, 2010; Munshi *et al.*, 2015). Clear measurable objectives are set for each simulation, and conclude with debriefing and reflection opportunities.

There is recognition that all acting is, by definition, 'not real'. The extent of the illusion is reflected in the different versions of reality and humanity portrayed, and in role-play it is possible to show both the inner and outer voices of a participant. This is referred to as the self-spectator, where participants are able to monitor their performance in the role, and are not overwhelmed by it (O'Neill, 2014). They are empowered through the initial setting up of the exercise to be able to maintain a dual personality: they are watching themselves as they play the role, and can learn from the

experience. This is particularly useful for researchers as it allows them to gather data from a dual perspective, i.e. what it was like for participants to play the role of someone else, and to compare that experience to participants' own realities. Such data can be used by the researcher to inform and create a multilayered approach to the research.

In the Stanford Prison Experiment, however, the participants appear to have been fully submerged in the role, or as Sartre (1976, p. 162) might put it, 'devoured by the imaginary'. This led to its own consequences, as presented in Box 30.2, and serves to highlight the necessity of thorough planning and preparation for the use of role-playing in research contexts, particularly in the field of social psychology. O'Neill (1995, p. 70) suggests that anyone who publicly takes on a fictional role changes in response to the alteration in interpretive attitude of the viewer to the viewed. They become simultaneously 'more' than themselves in a state of metaxis, acquiring what Roland Barthes (1972, p. 49) calls a 'corporeal exemplarity'. A role can protect and conceal a participant within the dramatic world, and they can be simultaneously 'both more and less than themselves. They embody both present meaning and future possibility' (O'Neill, 1995, p. 144). For this reason, role-play can provide an experimental setting in which questions of identity and the power and limitations of the roles we inhabit may be explored (O'Neill, 1995).

A further example of ethical dilemmas in role-play are the experiments by Stanley Milgram on obedience to authority (Milgram, 1974). In a series of studies from 1963 to 1974, Milgram carried out numerous variations on a basic obedience experiment which involved individuals acting, one at a time, as 'teachers' of a 'learner' (who was, in reality, a 'confederate' of the experimenter). The researcher explained to the participants that the experiment, though initially advertised as a study of memory, in fact was a study of the effects of punishment on learning: if 'learners' made errors in learning they were 'punished' by receiving what the 'teacher' believed were electric shocks. In fact, the teacher – the member of the public who was a participant in the research – did not know that the electric shocks were not real but were simulated by the learner (an actor who had been briefed by the researchers to complain, shout and scream as the supposed electric shocks were administered).

The experiment was not actually about the effects of punishment on learning at all; it was about obedience to authority, as the 'teachers' were urged strongly by the research director (an authority figure: a man dressed in a laboratory coat) to persist with the experiment even if they had serious reservations about administering increasingly powerful electric shocks.

'Teachers' were required to administer electric shocks of increasing severity every time the learner failed to make a correct response to a verbal learning task. Over the years, Milgram involved over 1,000 subjects in the experiment – subjects, incidentally, who were drawn from all walks of life (rather than, for example, from undergraduate psychology classes). Summarizing his findings, Milgram (1974) reported that typically some 67 per cent of his 'teachers' delivered the maximum electric shock to the learner despite the fact that such a degree of severity was clearly labelled as highly dangerous to the physical well-being of the person on the receiving end.

A role-play can involve scripted and/or unscripted elements. The 'confederates' might be given scripted comments in order to ensure some 'controls' in the role-play. For example, in this study the 'learner' was a confederate who gave scripted responses to different levels of electric shock, which were standardized across the different 'teachers', and the laboratory-coated director gave scripted responses to 'teachers' who were questioning the state of health of the 'learner' or who were reluctant to continue with the study.

Again, the research raises uncomfortable ethical questions for researchers using role-play. On the one hand, the research involved deception, telling lies about the purpose of the experiment, making the 'teachers' believe that they were administering electric shocks, making them feel very uncomfortable about administering these, putting them in a conflict situation and putting them under pressure to continue with the experiment despite their reservations, challenges and discomfort (Baumrind, 1964; Mixon, 1974).

Objections to the use of deception in experimental research are:

- Lying, cheating and deceiving contradict the norms that we typically try to apply in our everyday social interactions. The use of deception in the study of interpersonal relations is equally reprehensible. In a word, deception is unethical.
- The use of deception is epistemologically unsound because it rests upon the acceptance of a less than adequate model of the subject as a person. Deception studies generally try to exclude the human capacities of the subject for choice and self-presentation.
- The use of deception is methodologically unsound. Deception research depends upon a continuing supply of subjects who are naive to the intentions of the researchers. But word soon gets round and

potential subjects come to expect that they will be deceived. It is a fair guess that most subjects are suspicious and distrustful of psychological research despite the best intentions of deception researchers.

On the other hand, the participants were all volunteers. Moreover, they had the right to withdraw at any time or not to administer the electric shocks, they were debriefed at the end of the experiment so that they knew that no actual harm had been done to the 'learner', and neither Milgram nor any of the psychologists whom he had consulted before the experiment thought that the 'teachers' would persist in applying the electric shock of such high voltage, and indeed the participants indicated subsequently that, though they had been stressed, they were pleased to have taken part in the experiment (Dixon, 1987). Further, the experiment showed clearly the huge power of an authority figure to compel obedience – people simply follow orders and rules.

As with the Stanford Prison study, the Milgram role-play experiment, though it is unlikely that it would be approved by ethical regulation bodies nowadays, provided findings which might not have been possible otherwise. The dilemma for researchers using role-play again raises awkward questions: Is the benefit worth the cost? Does benefit to the many override the harm to a few? Does the end justify the means? Further, for researchers, the Milgram obedience to authority study demonstrates very clearly the power of role-play as a research technique.

30.6 Role-play as a research method

The approach to role-play advocated in this chapter represents a move away from a strictly behaviourist system to one which emphasizes process, and is involved in the creation and communication of meaning. Thus, role-play as presented here offers particular advantages to a researcher who is interested in exploring and analysing data which may not be easily accessed through other methods. It is a unique blending with case study method, and offers a rare opportunity to critically examine aspects of social behaviour and social interaction in relationships between people, ideas and the environment.

Using role-play in research allows researchers to:

- explore the principles of human behaviour in real-life settings, lived at life-pace;
- access and assess how people make sense of their lives, and the structures of the natural world;
- prioritize the process of engagement;
- explore different points of view, and forge different types of knowledge;
- adopt multiple viewing points within a data set;
- study multilevel communication;
- identify and explore the development and manifestation of participants' attitudes, decisions, strategies, values, higher-level cognitive and affective thinking skills, and emotions;
- shift and alter variables as the research unfolds, to explore subtleties and nuances in human interactions and situations, in an uncomplicated and undemanding manner, without having to schedule additional sessions or devise alternative methods to collect the data required;
- provide planned or spontaneous physical, emotional, personal, social or intellectual prompts and stimuli to participants, in comparison to the use of predominantly intellectual prompts in other methods such as interviews and questionnaires;
- engage with a fully diverse research population through the use of an inclusive method to explore and access relevant data;
- explore meanings and the ways in which people understand things;
- investigate patterns of behaviour;
- provide a somewhat objective lens through which to interpret the material, and thus distance themselves from the topic of inquiry, facilitating an objective mode of analysis;
- examine ready-made, visually and narratively rich research data, which evoke layers of meaning through reflection;
- capture visual data, adding immediacy and authenticity to the research;
- involve participants as co-researchers;
- engage in meaningful interaction with participants.

30.7 Role-play as a research method: special features

Role-play as a research method has several special features:

- Participants are actively involved in the research process through the three major stages of briefing, acting and debriefing. The use of role-play in a well-structured research process has the potential to create a reciprocal relationship, a valuable learning experience for both researcher and participant, which may ultimately impact upon the quality of resulting data.
- It helps participants to consider ideas from different perspectives. It can place participants at the centre

of the research experience, and allow them to construct their own bridges of understanding. As a result, they are often better able to respond to questions and comments from researchers about the experience.

- It supports participants during the research process, owing to the group and social nature of the activity. It tends to be much less isolating than completing a questionnaire, for example.
- Engages the whole person through the process, and reduces the danger of intellectual speculation or 'navel gazing' (what I would do if I were in that situation…). It places participants '*in situ*', at that moment, and demands a holistic response.
- The role-play can be structured to become incrementally more challenging or complex as participants are eased into the activity and prepared to engage with the issues under examination.
- It is an enjoyable activity and fosters positive relations between the researcher and participants.
- It is a spontaneous, dramatic, creative research strategy in which participants overtly and consciously assume the roles of others.
- The role-play may stimulate related memories and experiences, and can be used as a naturally occurring springboard to explore other relevant experiences or situations without the researcher probing too deeply or overtly.
- It can both relax and poise participants simultaneously, who may respond more openly and freely without overt direction from the researcher.
- It can be controlled by participants, and they can stop, pause or extend the activity at will.
- Debriefing and de-roling activities can increase reflection, and provide rich data that is not easily accessed using other methods, or within such an economic time frame.
- It can provide an added dimension to the research in that participants are engaged in reflexive praxis; they are learning and doing at the same time, i.e. research as a combination of both experience and reasoning.
- Like other forms of empirical data, role-playing may not provide researchers with unbiased, objective documentation, but it can show characteristic attributes that are often missed in other forms of data collection.

30.8 A note of caution

Like much research in the qualitative tradition, role-playing as a research method caters for issues concerning moral responsibility, individuality, freedom and choice, resulting in the collation of rich and personal data. While quantitative research is characterized by presupposed outcomes, qualitative analysis encourages an organic development, with much more flexibility offered to the overall process (see Bartlett and Vavrus, 2016). However, it is important to note that role-playing is always context-bound and localized. It does not lend itself easily to mass generalization, unlike quantitative techniques. The conclusions are usually derived from intensive, small-scale experiences drawing on a rich and deep data set, but they may be highly selective depending on the researcher's objectives. In this approach, the researcher turns away from statistical analysis in favour of in-depth analytical accounts of human behaviour.

If using role-play as a research method, the researcher must ask herself, what impact does the interplay of art with reality have? In addition to those issues identified in the previous section, Ginsburg (1978) summarizes the argument against role-playing as a device for generating scientific knowledge when he notes that:

- role-playing is unreal with respect to the variables under study in that the subject reports what she *would do*, and that is taken as though she did do it;
- the behaviour displayed is not spontaneous even in the more active forms of role-playing;
- the verbal reports in role-playing are very susceptible to artefactual influence such as social desirability; and
- role-playing procedures are not sensitive to complex interactions.

Ginsburg's (1978) critique relates to a form of practice that underpins a behaviourist approach to role-playing. It is noteworthy that many of his concerns have been addressed in the intervening years through the development of a systematic approach to role-play methodology as described in the following sections.

30.9 How does role-play work?

Maier *et al*. (1957, p. 14) state that leading role-playing is not a difficult task, and does not require special training by the trainer. However, this is disputed by Argyris (1958, p. 321), who claims that role-playing requires skilful leaders who have a high degree of self-awareness, confidence and self-worth. As alluded to earlier, the shift in role-play from a typical transactional model of passing on of knowledge to the 'making' of it, 'calls on one's humanness in a way not normally

associated with an instructional context': it can be both demanding and revealing (Bolton and Heathcote, 1999, p. 58).

This was evident in the response of the participants in the Stanford Prison Experiment, and the challenge for researchers is to get the balance right between maximizing opportunities for research and protecting participants. If one over- or under-protects, it may stifle learning. Many people are nervous about role-play, and associate it with being required to 'act' in front of their peers or colleagues. Taking on a role is like an actor working to create a character in a play or film. However, whereas the actor is required to build a complex personality through a process called characterization, the role-player focuses only on:

1. the purpose of taking on the role;
2. the status or level of power of the role – high, low or equal status in relation to the others in the role-play;
3. the attitude of the role; and
4. the participant's motivation in the role-play (O'Toole and Haseman, 1992, pp. 7–13).

The researcher should determine these in advance according to the questions or themes being investigated, and brief participants fully on these four points before they engage in the role-play (cf. Rao and Supans, 2012). This will facilitate transparency about the research exercise, and ensure greater clarity and depth in the activity itself, thereby improving the reliability of the data by more closely reflecting the real-life situation and reducing any tendency to superficiality. Using a variant of conversation analysis called the conversation analytic role-play method (CARM), Stokoe (2013, 2014) highlights the value of using animated audio and video recordings of real-time, actual encounters which differ from traditional role-played interactions, which do not always mimic and prepare participants for the real situations and events they are designed for.

Unscripted or improvised role-play increases flexibility, encourages varied discourse and allows for natural turn-taking in a conversational exchange, but educators and researchers should be aware that a loosely structured role-play places more demands on participants, and thus requires greater preparation in advance. A more structured, scripted role-play may also be used, but it does not allow for the same level of discourse and flexible response as improvised role-play. Occasionally, the educator or researcher may play a role (often called Teacher-in-Role in the literature), but usually most roles are assumed by participants. Cowley and Stuart (2015) report positive results of using a practitioner in role in their study of helping political science students understand the role of the whips in British parliamentary politics.

Although no set method or standardized approach exists for role-play, Waters (2016) suggests that where careful planning, implementation and management of the role-play is enacted, it helps to build learner confidence, enables deeper learning and assists graduate skills acquisition. The educator or researcher's ability to convey confidence to participants is key to ensuring a productive encounter. The eight principles outlined in Box 30.3 are designed to support the researcher as she endeavours to plan for a rich and well-designed role-playing episode. Without adhering to some general guidelines, a simulation or role-play is in danger of becoming stereotypical, overacted, simplistic and may skew resulting data. Such an activity may also peter out after a few moments if participants lack sufficient information about the situation or the characters they are playing. Depending on the researcher's objectives, it is possible to alter some of these principles below to elicit, monitor and assess a specific response. The following guidelines are useful to encourage active participation in most role-play situations, and should be attended to during the initial planning stage of the research, and communicated to the participants, either orally, much as a narrator in a film or play might do at the outset to fill in missing information, or through the use of written briefs or role-cards (which are commercially available or written by the researcher in accordance with her objectives).

30.10 Strategies for successful role-play

Inserting dramatic tension and awakening participants' self-spectator

Irrespective of the many types and genre of drama, one of its key defining characteristics is dramatic tension. Role-play, devoid of any tension, is sometimes used in educational and research contexts, and results in little more than rote learning or drill practice. While a behaviourist mode of training is appropriate in some areas, it can ignore the intricacies of real life where interactions with other people occur. If, as is suggested, dramatic tension is a key feature of successful role-play, then Heathcote (1991, p. 34) argues for the importance of focusing on the *quality* of dramatic tension, and by this she is not referring to 'huge terrifying events such as earthquakes, mutinies, armies and so on' which can characterize some forms of drama, but rather to localized incidents operating at a subtle level within a human circumstance. Tension is often manifest in situations where

BOX 30.3 MANAGING ROLE-PLAY EFFECTIVELY

1. *Set the scene*: when the participants are settled, the researcher should introduce the activity and outline what is going to happen during the session.
2. *Narrate the dramatic frame*: describe the context and background to the fictional situation by outlining any necessary information, i.e. what has happened up to this point in the story, where is this scene set, who is present, when does it take place, etc. It familiarizes the participant with the context, and removes some of the awkwardness associated with starting a role-play 'cold'.
3. *Provide a 'second dimension' for each role*: the researcher must provide adequate information about each of the characters in the role-play in order to 'flesh out' their profile sufficiently for the role-player to be able to 'step into the role' safely, confidently and with integrity. The 'first dimension' of role specifies only the character's broad profile, such as being 'a father', 'a teacher', 'a prisoner', 'a doctor', but does not indicate what kind of doctor is to be represented, what training she has had, what are her dominant personality traits (kind, generous, short-tempered, even-handed) etc. Talk about 'the character' as if you know her (it will increase participants' interest and investment in her situation).
4. *Dilemma*: the researcher must outline the dilemma or problem which is to be explored (and/or resolved) in the scene (usually consisting of conflicting choices where decisions have to be made and consequences dealt with). In planning the research, the dilemma or problem selected for inquiry may be of a personal nature (my family comes before my job), social nature (everyone goes to the nightclub at weekends), or of a moral nature (if we restructure the company in this manner, many workers will lose their jobs).
5. *Dramatic tension*: all drama, by virtue of its definition, relies on dramatic tension to propel the action forward. Tension may occur as follows: in relationships; as a result of a task that has to be undertaken; in not knowing what is going to happen (surprise and/or mystery); or in exploring ways of behaving not typical in participants' every daily lives. A successful role-play must have dramatic tension to sustain character belief and investment in the situation ('as if it could be real'). A well-chosen dilemma will lead to dramatic tension in the scene.
6. *Objective*: the researcher must ensure that *each* participant in the role-play has an objective. For example, as Human Resources Manager, invite the union representative to lunch, and your objective is to find out who is driving the proposed work stoppage among the workers. The person playing the part of the union worker may be given a different objective, possibly one that counters yours (i.e. reveal nothing), or operates at a more divisive or subtle level (provide misleading information, or play along with the game). Selecting the right objectives will impact on the focus of action in the scene, and thereby facilitate the researcher to gather data on his/her area(s) of interest. It will also impact upon the mood generated by the participants in response to their attempts to achieve their objectives, which may further alter or intensify the dramatic tension as a result.
7. *Constraint*: the researcher must formulate appropriate and purposeful constraints for each participant in the role-play before it begins. Constraints help to make a scene more realistic, and slow the action down, allowing for greater opportunity for negotiation and interaction. To be meaningful, constraints must be related to the dominant political, economic, historical, social or personal realities in the scene. For example, in the scene above, the union representative may be aware that he is being 'pressed' for information, but has to maintain a calm and vaguely pleasant demeanour as he is aware that greater harm could result if he were to have an outburst at a lunch table with the HR Manager. The constraint for the HR Manager could be that she is not allowed to ask the union rep directly about staff members' activities, and has to exercise caution in how she gently probes over a long and leisurely lunch. Without effective constraints, a role-player may ignore the social and professional 'niceties' in the scene above, demand the required information and conclude the scene rather swiftly, thereby missing out on the learning possibilities that this activity has to offer.
8. *Hidden objective*: while the information in principles 1–7 should be shared with all role-play participants openly, the researcher may decide to add additional information or instructions to complicate, enrich or develop a scene. It involves giving a piece of information or instruction to one participant in the role-play, and giving a different piece of information or instruction to the other role-player(s). This information is not shared with the full group but delivered privately, and thus when the characters come together to improvise

continued

continued

the scene, their objectives may clash overtly (or covertly) as set up by the researcher. A hidden objective can be used successfully to re-play the same scene, but altering some of the detail in the second and subsequent runnings. For example, in a re-run of a scene between a marriage guidance counsellor and a client, the researcher may call the counsellor to one side of the room to additionally inform her that this client has already seen another counsellor in the same organization, and made an official complaint about her. The client is not aware of this additional information being applied to the scene, and it would be interesting to gather data from both participants, comparing how the two scenes may (or may not) differ. It is possible to give a hidden objective to both parties, which can result in a lively interaction when the scene re-commences.

there is an incomplete task with a deadline looming, and related to power games and status in relationships. Inserting low-level or insipid dramatic tension into a role-play can motivate participants, build investment in the fictional situation and 'has the effect of making the most hackneyed situations spring into new focus and create new awareness' (Heathcote, 1991, p. 34).

Being cognizant of the fact that role-play operates in a fictional realm, and employs the art form of drama as its vehicle to achieve new insights, participants must not be allowed to become emotionally and intellectually consumed by a situation, or it will reduce the possibilities for reflection on their actions and related consequences. In addition to good planning for role-playing episodes, the concept of the 'self-spectator' (see Bolton and Heathcote, 1999; O'Neill, 2014) is closely linked with protecting participants in drama, and allowing for greater reflection and deepening of the experience. The concept implies that participants are observing themselves when in role, are aware of what they are doing and of what is happening to them, and do not become overly immersed in the action. This is achieved by monitoring their emotional and cognitive responses to the dramatic stimulus, so that they are aware that they are playing a part. They are simultaneously themselves and also representing a character.

Self-spectation implies becoming the critical audience of your own performance, and facilitates an ability to change if required. Failure to monitor one's participation in a role-play may result in missed learning opportunities, reduced flexibility in responding to a situation and increased risk of dangerous emotional engagement (i.e. getting carried away with the action). Regular moments of reflection both during and after the role-play are important to allow for self-spectation to be activated and employed. These can be facilitated by researcher interventions during the activity, such as questioning, judicious use of praise and encouraging participants to be responsible and to look for implications and consequences of their actions at all times. It is important to encourage the participants to document their experiences of the role-play, whilst in- and/or out-of-role. It can allow for the emergence of important insights and form the basis for later reflection and evaluation. Writing or drawing whilst inside or outside the dramatic situation can facilitate the formulation and expression of both private and public responses, and also further stimulate self-spectation.

Protection into role and protection into emotion

Emotion is the underlying currency of drama, because any imaginary act is necessarily accompanied by emotion (Davis, 2014). It fosters participant investment where the characters begin to care about the situation, and work collaboratively towards exploring creative and meaningful solutions. There are many ways to categorize and discuss emotion, but in educational drama and role-play we are broadly concerned with the notion of first- and second-order emotions. The former describes raw emotion as experienced in real life, and the latter refers to filtered emotion, as may be experienced in art (see Witkin, 1974; Best, 1992). It is widely agreed that first-order emotion has little or no place in art, as it is transitory and fleeting, and may at times be overwhelming and uncontrollable. But the advantage of using the arts in educational research is that they allow us to slow down time, pausing and dwelling a little on experiences that might otherwise be lost to us. This can be a useful approach in gathering valuable data for research. For example, when working with children who were prone to public release of inappropriate behaviours, such as tantrums or meltdowns, a role-play methodology was employed to investigate whether such children could learn to mediate and manage their emotional state using an experiential rather than a behaviourist intervention (O'Sullivan, 2016b).

In drama and role-play, the aim is not to protect participants *from* emotion, but *into* emotion, in order to maximize engagement and extend learning opportunities. There are several highly effective strategies, including the aforementioned self-spectatorship and second dimension of role, which serve to maintain and increase the objective distance between participants'

real lives and the fictional scenario they are working in. The challenge is to induct people comfortably and carefully into role to ensure that they are equipped to play that part safely and responsibly.

Whereas a professional actor will develop a whole technique to assist him in safely 'stepping into the role' of someone else (being careful to ensure a distance is maintained at all times between his real life and the 'role' he is playing), role-play facilitators and researchers who use this methodology typically just ask participants to 'be a…' (pensioner, sales rep, waiter, taxi-driver, prisoner) with little or no preparation for what it might mean to 'be a…', and little understanding of the consequences and associated responsibilities of asking someone to assume the role of another. Without adequate preparation, participants may have little or no option but to revert to a stereotype, as they have been given nothing else to work from. Thus, if asking young children 'to be pirates', they tend to rely on stereotyped images from film and television to base their representation on. The use of context and second dimension of role (and the principles discussed in Box 30.3) can considerably reduce this risk, and in this case encourage the children to explore what type of pirate they are role-playing. In building an initial profile, children can be encouraged to think about what they might have done (as a pirate) to be outlawed. Preliminary discussion to elicit information about what type of people pirates are, how and why they found themselves following that way of life, and what it means to be a pirate historically and in today's world, are effective strategies to:

a build belief and investment (i.e. 'I know about and care about this role');
b protect participants into taking on that role, leading subsequently to an emotional engagement with the role; and
c activate their self-spectator.

Attention to detail and thoughtful, responsible planning should always incorporate these strategies in order to simultaneously challenge and protect participants when engaged in role-playing. Supporting the idea that preparation and participant information are key to success in role-playing, Biziouras (2013) found that depending on the reading materials which his students were given before they undertook the role-playing simulation (i.e. different theories on international relations), different decisions and responses were evident in how the students behaved during the simulation activity.

Paying careful attention to how we induct people safely and responsibly into a role will elicit more reliable and ethical data. It would appear that while every effort was made to organize the research component in the Stanford Prison Experiment (via consent forms and university ethical approval), a major weakness in the design was evident in the lack of attention paid to the practicalities of using role-play as a research technique in this case study.

Box 30.4 sets out practical points to consider when setting up a multiple role-playing procedure.

The 'debriefing' stage of the role-play

A key element of a role-play for researchers is the debriefing. Debriefing involves sharing, discussing, reviewing and reflecting on experiences during the role-play, evaluating these and integrating them into the minds of the participants. The debriefing can be descriptive, evaluative, reflective and formative. It serves many purposes:

- to safely close a role-play and allow the participants to resume their normal roles and return 'to themselves';
- to review the contents of the role-play (what happened);
- to make sense of what happened (e.g. what were the key features);
- to share experiences and perspectives on the role-play (people's views and experiences will differ);
- to make meaning of what happened;
- to link the role-play to 'real life';
- to discuss and correct any errors in participants' knowledge, analysis and performance;
- to identify further learning development needs.

For the researchers, the main purpose of the debriefing is for participants and the researcher to learn. The comparison of role-play experiences of different groups can be part of the learning experience.

In this stage the role-play participants are also given feedback on what happened. For example, in the Milgram studies of obedience to authority, mentioned earlier, they were told what the experiment was for and that in fact no electric shocks had been administered to the 'learner', and steps were taken to ensure that they (the participants) suffered no after-effects or trauma. Further, it is in the debriefing stage that the researcher can also obtain further research data, for example, on the participants' feelings about, reactions to, views on, actions and behaviours in, thoughts about, reflections on what was happening in and what had happened in the role-play/simulation, the outcomes of the role-play/simulation, and so on.

Caution has to be exercised: too little debriefing and the role-play/simulation loses its purpose. Too much debriefing and it loses its impact. The researcher can plan the timing and duration of the debriefing carefully (a rule

BOX 30.4 PRACTICAL POINTS WHEN SETTING UP A MULTIPLE ROLE-PLAYING PROCEDURE

1. If the researcher is using 'a multiple role-playing procedure', where there are a number of pairs or groups conducting the role-play in the same space at the same time, begin by organizing the groups according to the number of people required in the scene (i.e. 'get into groups of three please').
2. Using the eight principles of role-play outlined in Box 30.3, give the group the requisite information and instructions for the ensuing role-play. Ask them to take a moment to discuss who is going to be whom and what the characters' names are (if this has not already been predetermined). It is a good idea to write the names of the characters on a clearly visible flip chart which participants can refer to if they forget, without stopping the action.
3. Ask the participants to find a space in the room and organize it in preparation for the role-play (i.e. loosely demarcate it as an office space, a shop, a university classroom, etc. according to easily identifiable and available objects and resources). Invite them to use a chair or bag to section off their space. This helps to establish belief in what they are doing and makes their space semi-private so that they can focus on the task in hand.
4. Inform them that the role-play will begin at the same time for all groups, and that when theirs has run its natural course, they should remain *in situ*, and quietly observe until the other groups have finished. On average, role-plays will run for between five and ten minutes if they have been well set up.
5. Where the role-play begins with all players *in situ* (i.e. sitting opposite each other in an office-type setting), the researcher should invite them to adopt an appropriate position for their role, such as scanning through a list on the desk, fiddling with a watch to signal nervousness, or concluding a phone call. Ask them to place their eye contact on an object rather than on their partner(s), freeze this gaze and their physical action for a moment, and on a clearly audible count of 1, 2, 3 from the researcher, all groups begin at the same time. It is a good idea to provide the first line or opening words of the scene, such as, 'Now then, Mr Hayes, why did you come to see us today?' which all groups can recite to get them started. It reduces tension and any nervousness that may be present, and can usefully serve to focus the direction, the tone and the mood of the role-play.
6. If the role-play begins with one person entering a room and the other(s) already *in situ*, ask the person entering to stand about a metre away from their role-play partner(s), to lower their gaze as if preparing to knock on a door and enter on a given signal by you. In this situation, the person/people inside the room should adopt an appropriate action and eye gaze, away from the imaginary door and their role-play partner standing 'outside', and wait for the facilitator to count aloud and knock physically on a table or wall on everyone's behalf. The opening words provided here may be something like: 'Come in please', or 'One moment please, I'm on the phone/finishing off a document', etc. Deciding to leave a person waiting outside an office door for 10–20 seconds can create an interesting power dynamic that will impact upon the remainder of the role-play as it unfolds.
7. Allow the role-play to run its natural course, and if most pairs/groups are finished, you can gently intervene by inviting those still going to finish up shortly.
8. Reminding the participants that they are 'back to themselves', provide an opportunity for the pair/group to reflect and discuss initially, and then open up the discussion as a whole group exercise. This will be structured according to the individual research requirements. Participant diaries can be a useful tool to gather participant perspectives.
9. Scenes can be re-played as required, giving participants a different experience by shifting or altering any of the principles outlined in Box 30.3.

of thumb is that a debriefing may take up a quarter of the total role-play/simulation time at most, and much less if it (e.g. a simulation) has run over several sessions).

The researcher will need to decide whether to conduct the debriefing in a plenary session and/or with individuals or groups (the latter can be time-consuming).

If the debriefing takes place immediately after the end of the role-play (which is a good idea), it is advisable to have a few minutes break, so that the debriefing clearly marks a new stage of the role-play.

It is useful to consider the physical layout of the space for debriefing: a circle or semicircle/horseshoe is desirable as everyone can see everyone else, and there is little issue of 'status': all are equals, with the Chair simply managing the discussion or interjecting with feedback and/or questions.

It is essential for the researcher to prepare the debriefing, as the debriefing is a guided discussion, not a 'free-for-all'. The debriefing must strive to be positive, with the Chair showing empathy rather than being judgemental (unless there is a point of correction to be made). Sensitive and productive debriefing is a skill, an art, requiring an ability to tolerate ambiguity and difference.

The researcher must consider many questions in preparing for the debriefing:

- what are the purposes of the debriefing (and tell the participants);
- what points *must* be addressed;
- what questions can be posed (e.g. to promote discussion and feedback);
- what feedback must, should and may be given, and about what;
- what data the researchers wishes to collect;
- what and how to summarize;
- how to involve and receive feedback from observers (if there were any);
- how to manage the debriefing:
  - what are the ground rules, for example: the Chair nominates or identifies the speaker (or places several volunteer speakers in a sequence)
  - nobody interrupts a speaker
  - speakers are polite and respectful, even when raising points of criticism
  - nobody may speak for more than three or four minutes
  - each speaker must keep to the point, i.e. avoid redundant or irrelevant material
  - each speaker should build on, or link to, the previous speaker where possible;
- how to affirm participants and be positive about comments received;
- how to promote participation in the debriefing;
- how and when to give feedback from observers/experts.

Opening the debriefing session

The Chair indicates what will happen in the debriefing, what are its purposes, how it will be managed, the importance of keeping to the point, and that there may be no right answer or no single solution to issues being raised. Then the following can take place:

- Invite participants to volunteer their views on how they feel about the role-play (a general question to see if this promotes a response to 'break the ice').
- Invite participants to volunteer their views on what happened, what they did and why, what they felt went well and less well, and why.
- Invite participants to comment on their roles, interactions and reactions to others, how effectively they operated in these and why/why not.
- Invite participants to share their *feelings* about the role-play, for example, which parts they enjoyed, which parts excited them (or the opposite), what surprised them, and why, and what were their feelings. Gain reactions of others to the same point/situation being made.
- Which parts did the participants find challenging, difficult, easy, and why?
- Where did they feel they were most/least effective, and why?
- What would they do differently next time, and why?

Continuing the debriefing

- Invite participants to identify *key features* of the role-play and what they *learned* from it (e.g. the topic, themselves, the interactions, the relationships, their roles, conducting a role-play).
- Why did people (themselves and others) behave in the way that they did? Are there other interpretations of what happened?
- In what ways was the experience of the role-play beneficial and worthwhile?
- What aspects do they need to learn more about?

Concluding the debriefing

- Receive feedback from observers/experts.
- How to apply their learning in the 'real world'; what were the key points of application.
- End the debriefing on a positive note.

For some reticent participants, the Chair must judge whether to invite their direct participation.

30.11 Examples of research using role-play

As noted earlier in the chapter, role-play appears to be making a strong comeback in the broad field of education and training. It is no longer being given a cursory mention in research studies, but is beginning to feature as a key teaching, learning and assessment strategy in many publications across a diverse range of academic and professional disciplines. While there is still more work to be done in terms of encouraging researchers to consider its use as a valid and valuable research method, Table 30.1 provides a brief summary (illustrative only) of some recent reports on role-playing, both as a research technique and as an effective approach to teaching and learning.

TABLE 30.1 EXAMPLES OF THE USE OF ROLE-PLAY IN THE LITERATURE

Article	*Use of role-play*	*Major findings*	*Issues for consideration*
Seiler, S. N., Brummel, B. J., Anderson, K. L., Kim, K. J., Wee, S., Gunsalus, C. K. and Loui, M. C. (2011) Outcomes assessment of role-play scenarios for teaching responsible conduct of research. *Accountability in Research*, 18 (4), pp. 217–46.	As a form of summative assessment to teach responsible conduct of research (RCR) to graduate students in science and engineering.	Results suggest that role-playing might promote a deeper appreciation of RCR by shifting the focus away from wanting to simply 'know the rules'.	The authors also used a *think-aloud* case analysis approach to assess participants' case analysis performance.
Johansson, J., Skeff, K. M. and Stratos, G. A. (2012) A randomised control study of role-play in a faculty development programme. *Medical Teacher*, 34 (2), pp. 123–8.	To investigate the impact of role-playing as an instructional technique for facilitating change in teaching behaviours.	Data from 48 hospital physicians indicated significantly greater positive changes in teaching behaviour among faculty who attended the standard course (with role-play) as compared to those in the alternative course ($p = 0.015$).	This study validates a commonly held view in health sciences education that role-play is a useful instructional method for improving teaching.
Stevens, R. (2015) Role-play and student engagement: reflections from the classroom. *Teaching in Higher Education*, 20 (5), pp. 481–92.	To elicit feedback from 144 history students about their experience of a role-play activity, identifying what they gained from the activity and if it encouraged them to learn more about the topic.	A large majority found the role-play activity beneficial, but a small minority reported gaining little from the exercise. The author argues that role-play may be less beneficial for weak or unprepared students.	It is acknowledged that although role-play may be a popular teaching method, the manner in which the teacher/facilitator sets it up and prepares the students can impact upon learning outcomes.
McEwen, L., Stokes, A., Crowley, K. and Roberts, C. (2014) Using role-play for expert science communication with professional stakeholders in flood risk management. *Journal of Geography in Higher Education*, 38 (2), pp. 277–300.	To explore role-play pedagogies and evaluate participant perceptions of their learning experiences in learning and communicating about flood science by flood risk management professionals in local government.	Results suggested the development of analytical and strategic use of flood science skills, and increased confidence in science communication.	The negative impact of prior role-play experiences affected participants' attitudes to learning, and learner diversity affected co-learning.
Browning, T. R. (2014) A role-playing game for teaching about enterprise process integration. *Journal of Enterprise Transformation*, 4 (3), pp. 226–50.	To examine the impact of a role-playing game that helps teach process integration in a more streamlined, customer-oriented manner.	Playing the roles of owners of various enterprise processes who must coordinate their input–output relationships, participants' awareness was raised to the challenges and potential of integration at the enterprise level.	The role-playing game led to fruitful discussion and negotiation, to successfully bridge the gap between poor understanding of the motivations and methods for successful integration.

Article	*Use of role-play*	*Major findings*	*Issues for consideration*
Carte, L. and Torres, R. M. (2014) Role-playing: a feminist-geopolitical analysis of the everyday workings of the Mexican state. *Gender, Place & Culture*, 21 (10), pp. 1267–84.	Examining in detail the implementation of role-play as a research method in a case study with Central American immigrants.	Role-play was shown to be very useful in revealing immigrant daily experiences as they try to assert their rights.	The authors report that role-play is particularly suited to revealing immigrant women's experiences due to its encouragement of creativity and facilitation of discussion around challenging subject matter.
Guilfoyle, N. and Mistry, M. (2013) How effective is role-play in supporting speaking and listening for pupils with English as an additional language in the Foundation Stage? *Education 3–13*, 41 (1), pp. 63–70.	The study investigated how role-play supports the development of language skills for young EAL learners.	The authors found that role-play promoted the use of a wide range of key strategies for language learning.	Role-play is well adapted for use in early years language pedagogies.
Deaton, C. C. M. and Cook, C. (2012) Using role-play and case study to promote student research on Environmental Science. *Science Activities: Classroom Projects and Curriculum Ideas*, 49 (39), pp. 71–6.	To investigate the use of an integrated role-play and case study approach in developing critical thinking skills, communication skills and learning communities.	Students were actively engaged with a scientific issue through a case study approach, and took on the role of a case study character to research environmental science.	This integrated approach successfully combined the gaming features of role-play with the narrative of case study.
Maratou, V., Chatzidaki, E. and Xenos, M. (2016) Enhance learning on software project management through a role-play game in a virtual world. *Interactive Learning Environments*, 24 (4), pp. 897–915.	Using an immersive multi-user virtual world, the role-play game aims to enhance experiential learning of human-related issues such as communication and collaboration with other team members, which the authors indicate are not easy to teach through other methods.	Participants evaluated the game positively in terms of overall game experience, enjoyment and learning impact, commenting on the successful challenge of interacting with other online players.	The instructor is able to observe the players, intervene when needed, and significantly alter specific game scenario parameters to modify the level of challenge and difficulty for players.

30.12 A note on simulations

The difference between role-play and simulations

Simulation is a commonly used term in research, and increasingly is used almost interchangeably with role-play. However, traditionally, to simulate meant to imitate, to pretend, to copy, and was typically associated with a person 'pretending' to be in a different situation (often as themselves). It was regarded as a powerful form of active learning and engagement, allowing participants to experience and possibly rehearse/practise other situations within the domain of 'real life'. Simulations are often used in games and training scenarios, such as for medical personnel, pilots, engineers, construction workers and military personnel.

Simulations differ from role-play in several key respects, but principally in terms of the degree of engagement with the role assumed. Role-play can be more demanding and generally requires greater imagination and preparation on the part of the role-player to successfully 'get into' the role and 'take on' more of the depth, complexities, subtleties, nuances and challenges associated with temporarily 'being' that person, in that situation. Simulations require participants to take decisions in response to their assessment of the situation they have been placed in, and to evaluate and monitor their performance and its impact on others, while role-play is more exploratory and may not lead to decisions but to further exploration of problems in a spontaneous enactment.

In recent years, there appears to have been somewhat of a merger between the two ideas, and researchers often refer to role-playing simulations, where they have combined the elements of both concepts to create an interactive and dynamic approach. The use of online role-play simulations in particular has burgeoned in the last five years and is proving effective as a training intervention and research strategy in such areas as suicide prevention using emotionally responsive avatars (Bartgis and Albright, 2016), and developing transnational global competencies for engineering students entering an international workforce (May *et al*., 2014).

Simulations

Educational simulations involve a sequence of events which typically involve or lead to decision making, and in which the environment or situation is set up by the researcher (Hertel and Millis, 2002, p. 15; Cheng *et al*., 2014). In these, participants take on roles and the sequence of activities evolves, sometimes planned and sometimes unplanned, as the simulation rolls out over time. They often involve problems and many differences between participants, for example, in terms of attitudes, agendas, perceptions, powers, voice, status, values, sympathies and agendas, i.e. in which the potential for conflict or disagreement features strongly. Simulations are designed to put participants in a realistic representation of a situation, environment and issue, recognizing and building in the complexities of these, within which interactions should occur, leading to an outcome (e.g. a decision, a compromise, an action, a statement, a change of view) (Shaw, 2010, p. 2).

In setting up the simulation, researchers must ensure that the scenario design has the hallmarks of realism (e.g. situational, physical, emotional, conceptual, interpersonal), i.e. that it could really happen in the 'real world', and that participants understand the situation and what is required of them (Cheng *et al*., 2014).

It is important for the researchers to choose a topic that is interesting, topical, real-world, relevant and engaging (cf. Livingstone, 1999; Goedert and Rokooei, 2016). In choosing a topic, particularly for a simulation, it is often useful to select one that will enable different perspectives, interest groups (maybe conflicting interests) and agendas to be included, and this will enable different groups of participants to take on the roles/interests/agendas of these different groups. It may be useful to take a controversial issue, an issue on which there is dissensus and maybe strong feeling, a conflict or problematic situation, one that is not susceptible to simple solutions, or one in which there are different interests at play. It may be helpful to identify an issue on which there are many perspectives and differences of opinion, and then construct a scenario around this. Or the role-play/simulation may focus on a crisis situation, or a sensitive, delicate matter, or a situation in which there is no single or correct answer. By choosing a current issue for a role-play/simulation, this gives some immediate relevance to the topic.

Simulations:

- are 'real-world' and focus on 'real-world' issues, but in a safe learning environment;
- are based on reality, and focus on those parts of that reality which are deemed to be relevant for the case in hand; they are representations of reality;
- simplify a complex reality to focus on key issues;
- are participatory;
- hand over significant autonomy, responsibility and power to participants;
- use active, interactive and collaborative methodologies, promoting collaboration as well as competition;
- focus on the processes as well as outcomes;

- develop interpersonal, team-working and leadership skills;
- develop decision-making skills and include opportunities for prompt feedback to be given and received;
- focus on key issues and 'design out' distracting and extraneous matters;
- accelerate and condense certain events;
- develop negotiation and bargaining skills;
- develop abilities to look at a situation from many perspectives.

They typically have several defining features:

- key objectives are stated explicitly;
- simulations are carefully structured and timed, and are based on the key objectives;
- participants take on roles and act in role in the given setting/context of the simulation;
- there are interactions between the participants in role;
- the interactions are rule-governed;
- the outcomes of the simulation follow from the simulation itself and its participants;
- success criteria for the achievement of the objectives are made explicit;
- there are often no simple or single solutions to the problems set out in the simulation.

Simulations have the potential to create complex, dynamic and evolving political processes and interactions, thereby enabling participants (including the researcher) to investigate and examine participants' (individually, by group, party etc.) motivations, constraints on behaviour, attitudes and values, and interactions among the actors in the situation.

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Visual media in educational research

CHAPTER 31

The chapter suggests an agenda for attention in using visual media and what practical considerations researchers can address in this enterprise: who provides the image; photo-elicitation and how to work with it; strengths and weaknesses of different kinds of visual media for research; how to interpret artefacts and images (and the issue of multiple interpretations); the importance of reflexivity. It suggests that visual data, whilst being useful as free-standing data sources, are also useful in conjunction with other kinds of data.

We are surrounded by visual data; knowledge comes in various forms and is not reducible to language alone (Eisner, 2008, p. 5). Visual methods can provide data that word-based data cannot (Clark *et al.*, 2010, p. 86). Visual data are increasing in educational research (e.g. Wall *et al.*, 2012) and this chapter introduces how researchers can use visual data, addressing a core of issues in the planning and conduct of data collection using visual media of different types. The chapter raises a series of issues concerning:

- photographs and still images
- video and moving images
- artefacts
- ethical practices in visual research

This chapter should also be read in conjunction with Chapter 36 on the analysis of visual data.

31.1 Introduction

Educational researchers can draw on a host of visual media in their research. These include, but are not limited to: film, video, photographs, television, advertisements, pictures, artefacts, objects of fine art, memorabilia, advertisements, moving images, still images, media images, maps, graphs, drawings and sketches, illustrations, graphical representations, cartoons, artefacts and everyday objects, deliberately non-commonplace objects, family photographs, and so on. In short, anything we see, watch or look at counts as a visual image. They are the stuff of ethnographic and anthropological educational research (witness, for example, the attention given to artefacts and visual images in studying organizational culture, and the messages about the organization that are conveyed in such images, discussed later in this chapter).

Using visual media concerns the production of the image, the image itself and the audiences of the image. Visual media are not neutral; they give messages, deliberately or not, and we interpret them in many different ways. They have their own forms and effects (e.g. compositions and technical properties, and these have an effect on the viewer). They are constructions of social events and perspectives, of power and power relations, of social relations and social difference. More than this, we look at them in different ways, i.e. we bring our own values, biographies, cultures and background to bear on images (Rose, 2007, p. 11). Images, then, must be viewed in the social and cultural contexts of their production (Banks, 1995, p. 2) with consideration of who are the audiences, intended or otherwise, of the image. An essential feature of an image is its audience and the way in which the audience views and 'reads' the image (Fiske, 1995). As Berger (1972) made clear, we have different 'ways of seeing'.

Images are made, kept and displayed in different places, from museums, cinemas and galleries to each person's home, each of which confers its own required social behaviours and audience reactions (as Bourdieu and Darbel (1991) indicated: middle-class, educated visitors to art galleries stand in quiet contemplation of paintings). Some visual media have texts, others do not.

A constant feature is the subjectivity of the producer and selector of the images. A professional photographer can use images to persuade, to project an agenda. A documentary maker can select video clips to press a point, and this becomes an ethical issue. A researcher can decide which photographs to use in a photo-elicitation technique, or how to brief children-as-photographers in creating their own images. The issue of bias has a high profile in image-based research, and the researcher's reflexivity is a key issue here.

An image is the product of *technologies* (oil paintings, video production, photographic materials, computer software), *compositional features* (visual form,

material form, presentational form, structure, colour (e.g. hue, saturation, lightness/darkness), texture, abstraction, expressive content, spatial arrangement, symbolism etc.) and *social contexts* (cf. Rose, 2007, p. 26). Some images balance colour and content harmoniously; others scream at us. Some are close up, some are distant or wide-angle. Some shots are deliberately taken from an elevated, low, side or frontal position; some are posed; others are snapshots taken as the opportunity arises. Some are in focus; others are not. Some are geometrically structured (e.g. with perspective); others are free of geometric form. Some images are meticulously planned; others are fleeting snapshots taken on the spur of the moment. Some are part of a series or a collection; some stand alone. Some are part of a recognizable genre; others are not. Some are made by amateurs; others by professionals. Some are deliberately designed to give messages; others are not. Some are reflections of culture and society; others are in the vanguard of social and cultural change. Some are part of normal living (e.g. food); others are deliberate constructions that are out of the ordinary (the oil painting). Some are faded and fuzzy (the 'materiality' of the image (Rose, 2007, p. 234)); others are crisp and sharp. As Rose (2007, p. 26) remarks, visual images are never innocent; they are wrapped up in many layers of meaning and interpretation. They are not only 'reproductions of reality' (Flick, 2009, p. 240) but, rather, 'presentations of reality' which themselves are then interpreted by viewers. All of this renders images difficult to interpret and capable of multiple interpretations.

Huge proportions of the population can take still and moving images with conventional and video cameras, with both of these on their cellphone. Cameras can present an immediate, comprehensive and holistic image of situations, objects, people, events, lifestyles, contexts, conditions and so on, that happen very quickly or suddenly (maybe too quickly or with too many details or with too great a level of complexity for conventional observational recording to be able to catch). Such images are easy to transport, and enable the researcher to review them repeatedly (particularly useful for fleeting, short-lived and ephemeral moments) and indeed to have their reviews checked by a third party. Further, images can be taken non-intrusively, reducing observer effects and reactivity (cf. Denzin, 1989, p. 203; Flick, 2009, p. 241).

It may be the researcher who takes the image, or researchers can ask participants to take images, maybe even providing them with the camera so that they can decide what they consider to be important to be kept as a still or moving image (Flick, 2009, p. 242). Research that uses images may be both collaborative and participatory in involving participants as partners in the creation, production and discussion of images, both still and moving (but setting up a concealed camera is far from collaborative (Banks, 1995, p. 3)). As online communication increases, so the ability to share images has become part of everyday life. Indeed Banks (1995) argues that the dichotomy between the researcher and participants, the observed and the observer, is collapsing (p. 1).

The researcher can provide the already-taken images (and use them as a starting point for discussion), or ask the participants to bring images that they have and which they have or have not taken themselves (e.g. family photographs), and which can be used as starting points for interviews or as main elements of interviews (the 'photo-elicitation interview' (Harper, 2000, p. 725)). Here consideration has to be given to the taking of the image and the derivation of data from the image (cf. Denzin, 1989, p. 210).

In considering visual images, Denzin (1989, pp. 213–14) indicates that whilst cameras report what they see and what really happens (rather than the selective observation of the human observer), nevertheless images are selective, in that the image maker has already decided what to include or not, what to focus on and what not to focus on (Becker, 1986, pp. 241–2), where to point the camera and where not to point the camera. Images create their own representational and symbolic forms and they are time-bound; they catch a particular moment (or several). Given this, it is wise to regard visual images as telling a story – a discourse – rather than being a singular objective reality. Indeed it is commonplace to have written text – a commentary or analysis – accompanying the image, and this text, too, tells a selective story or has a selective focus.

Visual data catch and store a wealth of data in a single image or video sequence and, like other forms of observational data, they are selective in their focus and contents (e.g. deriving from the researcher's agenda, interests, research questions etc.). This presents issues of data overload, selectivity and manageability. Whilst this may present problems in the stage of data analysis (the problem of bias in being unavoidably selective), visual data can be one of a range of different types of data (e.g. written, aural, oral, observational) in a research project. Rather than standing on their own, visual data are one element in triangulated data and, as we mention in Chapter 36, can employ analytical techniques used with other kinds of data, as well having their own methods of analysis.

In contemplating images the researcher has to consider how much they are natural, contrived, arranged, posed or staged. In this respect there is an argument for

covert research and/or a fixed camera as it leaves the natural situation undisturbed.

31.2 Who provides the images?

Prosser and Loxley (2008) identify four main kinds of visual data: found data; researcher-created data; respondent-created data; and 'representations'. At issue here is who provides the data, for example, who are the photographers, who selects the photographs or edits the video?

Whilst it is commonplace for the researcher to provide the image (Prosser and Burke, 2011), there is a growing recognition that, in educational research, asking children to provide the images – either images they have taken or created themselves, or that they have collected or brought (e.g. family photographs, Internet images, magazine images, collages) – is a way of empowering participants and building rapport, not least in situations where, for example in researching with children, there are differences of age, status, power and language and explanatory abilities (cf. Hatten *et al*., 2013; Torre and Murphy, 2015). Researchers and participants become equal partners in the co-construction of meanings (Torre and Murphy, 2015, p. 6). In the case of participant-provided images, the researcher will need to brief them and provide an agenda and guidelines. Smartphones enable children to take still or moving images easily; they are digital natives.

Further, enabling children to create their own images is a doorway into their worlds and cultures in their terms, which researchers otherwise might not be able to enter or understand. Visual images enable children to present their worlds in *their* terms, meanings and perspectives. Children's worlds are highly visual and central to researchers' understandings of childhood; indeed, Harper (2002, p. 13) suggests that, in human evolution, visual processing evolved before verbal processing and is located more deeply in the human brain, evoking 'deeper elements of human consciousness than do words'.

31.3 Photo-elicitation

Photo-elicitation, a term penned by Collier (1957) when researching environmental influences on stress and mental health, has a history reaching back to anthropological studies in the late nineteenth and early twentieth centuries (Hatten *et al*., 2013; Torre and Murphy, 2015). Elicitation concerns disclosure of the 'core definitions of the self' (Harper, 2002, p. 13), 'ideas that they don't usually talk about' (Barton, 2015, p. 179), 'surfacing the taken-for-granted ideas about the social world' or other aspects of lives which might be easily overlooked (p. 197).

There are several types of photo-elicitation (Hurworth, 2003; Karlsson, 2012; Mills and Morton, 2013; Mitchell, 2012; Barton, 2015; Elliot *et al*., 2016), for example:

- *autodriving*: the researcher and/or participants can provide the photographs but participants take the lead;
- *reflexive photography*: participants take the photographs and then, at interview, are asked to reflect on these;
- *photo novella*: participants take photographs which tell a story about part/all of their lives;
- *photovoice*: participants photograph those parts of their society/community/environment (widely defined) which have meaning for them or which they feel need to change (i.e. the appeal to critical research);
- *photo-observation*: photographing real objects, events, activities and discussing them;
- *photo-interviewing*.

Photographs enable people at interview to talk about issues that word-based approaches alone cannot do; they stimulate discussion, provoke a response and enable the researcher to work collaboratively with participants (Richard and Lahman, 2015; Elliot *et al*., 2016). They can elicit information which researchers cannot directly observe: participants' ideas, emotions, inner feelings, perspectives, opinions, meanings and ways of thinking (Richard and Lahman, 2015; Elliot *et al*., 2016). Photographs may be the starting point of an interview (an 'ice-breaker') (Prosser and Burke, 2011), a supporting part or means of opening up an issue. They make concrete those abstract concepts or issues that are difficult to put into exact words, particularly for participants (e.g. children) who may not have sufficient linguistic or abstract thinking skills, or who may find it difficult or uncomfortable to talk about issues, or who need a safe environment in which to disclose information or intense feelings (Pyle, 2012; Hatten *et al*., 2013; Richard and Lahman, 2015). Photographs can trigger feelings (suppressed or forgotten), recollections, understandings, attitudes and opinions; they are evocative and often ambiguous and polysemic, and this can stimulate discussion and reflexivity (Harper, 2002; Pyle, 2012). Indeed Barton (2015) suggests that participants may find it harder to lie about their reactions to photographs because photographs have 'emotional salience' (p. 197).

Photographs, taken by the researcher, participants or other parties and brought to an interview, help

participants to focus, clarify, illustrate and explain an issue, to judge the importance given to the issue or the personal meaning and significance attached to it by participants. They help to elicit tacit or abstract knowledge that participants may be unable to or unwilling to share; they may bring an emotional dimension to the interview or matters in hand; they may facilitate conversation between relative strangers (researcher and participant(s)) and they give power to the participants (Meo, 2010; Hatten *et al*., 2013), bridging the gap between researchers, educators and students (Torre and Murphy, 2015). Photo-elicitation empowers participants and can build rapport and trust between them and the researcher, enabling the researcher to see the world and the situation through the eyes of the participants, and thereby give greater validity to the data and the participants' responses (Meo, 2010; Pyle, 2012; Elliot *et al*., 2016).

Photographs carry meanings that words alone, spoken or written, cannot. They convey real life, flesh and blood (witness Sutcliffe's nineteenth-century photographs of the fishing port of Whitby in the UK and everyday lives that he photographed, or the photographic work of Forsyth in the poor districts of twentieth-century Newcastle-upon-Tyne in the UK). Photographs evoke meanings and reflections as well as information and factual data. They catch the texture, the mood, the atmosphere, the 'feel' of real life and different places, emotions and flesh-and-blood drama. They frame how we think; they can mirror our thoughts or stimulate them. They are both *emic* and *etic*. They carry documentary and interpretive meanings, posed or natural. They can support and supplement other sources of data and text, or they can stand alone. They are less time-consuming to study than film footage or video materials. Indeed they are time-efficient and researcher-efficient, as they can convey far more in a single image than many pages of text ('a picture paints a thousand words').

A photograph can be of a real situation, an image – concrete or abstract, contextualized or decontextualized – which is a metaphor that is subsequently interpreted at interview, a representation of something, a symbol, a real event or an interpretation etc. (Richard and Lahman, 2015; Elliot *et al*., 2016).

In using photographs, researchers can take photographs and ask the participants to comment on them, or the researcher can ask participants either to take their own photographs (and the researcher might supply the camera) or to bring along to an interview (e.g. individual or group) one or more photographs that have meaning to them, to discuss them or provide a commentary on them. Such interviews or textual material can then use conventional methods of data analysis, for example, analysis of transcripts, field notes, software packages, coding, content analysis, grounded theory approaches, constant comparison of images and codes, looking for patterns and genre, and moving towards generalization where appropriate (see Chapters 32 to 37).

A seven-step sequence can be adopted in photo-elicitation (cf. Torre and Murphy, 2015):

Step 1: The researcher sets the topic for the research or investigation.

Step 2: The researcher identifies and invites suitable participants for the study.

Step 3: The researcher briefs the participants about the purpose, agenda, requirements, operation, ethics, constraints and conduct of the photograph provision, for example:

- the purposes of the photographs;
- what to do and why;
- what to focus on and why;
- who takes/collects the photographs (and individually or in pairs/groups);
- how to take photographs (if, for example, children are unfamiliar with how to handle equipment carefully and take photographs);
- what to photograph, not to photograph, and why;
- when to photograph (e.g. time of day, how frequently, at what intervals);
- where to photograph and not to photograph;
- to whom to show and not show the photographs;
- ethical issues of identification and anonymity, privacy, confidentiality;
- how many photographs (e.g. a maximum or how many to select);
- how to proceed (which might include children being accompanied by an adult);
- how the photographs will be used at the interviews.

Step 4: The researcher and/or participants decide who will take and/or collect the photographs.

Step 5: The researcher and/or the participants take and/or collect the photographs.

Step 6: The photographs are brought to, and form part of, the interview or discussion (see Chapter 25 on interviews – and whether they are individual or group – and their conduct).

Step 7: Data are analysed and the results reported.

Some of these steps may be in a different order or recursive (e.g. Steps 3 and 4) and, indeed, researchers

and participants may negotiate and agree the 'rules of the game' rather than the researcher taking all the decisions alone.

In the photo-elicitation technique, the photograph, or set of photographs, or sequence of photographs, is used to invoke, prompt and promote discussion, reflections, comments, observations and memories (Banks, 2007, p. 65). The interview or meeting between the researcher and participant(s) can start with photographs, what they show, who took them, when, where, what is the story behind them, and so on. Photographs can break down differentials of power between the researcher and participants (Prosser and Burke, 2011; Pyle, 2012; Torre and Murphy, 2015).

An image can mean different things to different people, there is no one 'correct' or meaningful interpretation of what it says about the world. A photograph of a child sitting at a desk can be seen as, for example: hard work; commitment; punishment; loneliness; struggle to learn; boredom; enjoyment of solitude; examination pressure; motivation; an outdated pedagogical strategy; delight in reading etc. The researcher has to ascertain what the image means for the participant, what significance it has, i.e. to respect the 'positionality' of the participants (Hatten *et al*., 2013) (see also Chapter 15).

Using photographs in an interview can overcome awkward silences or maintaining direct eye contact (Banks. 2007, p. 66), as this can be intimidating for some participants (e.g. children), not least because of the potential power and status differentials between the researcher and participants. The potential discomfort of face-to-face contact is alleviated by shared face-to-photograph contact. Further, having a focus on a photograph or different photographs can offset any feelings that the interview is some kind of 'test' or 'grilling' for the participants (p. 65), particularly if the photograph comes from, or has been taken by, the participant(s). Having a common/shared focus in the photograph introduces a 'neutral' third party (the photograph) into the interview (p. 66).

During the interview in which the photographs are used, participants may be asked to select some photographs and explain why they chose them, what the photograph(s) is/are about (different meanings and perspectives). They may be asked to sort and group photographs and explain their grouping criteria, or to arrange them in an order or along a continuum, for example, what they like best to least, what is most/least like them, what is closest or truest to their own lives etc. Such sorting can elicit the participants' conceptual categories: how they think and group items, and why (Barton, 2015). As in other interviews, the researcher must be ready with prompts and probes (see Chapter 25), helping participants to explain and crystallize their and the photographer's thoughts, meanings, perspectives, feelings and psychological states.

Whilst the researcher can strive to have high-quality photographs and reproduction, this is not always possible: old photographs fade over time; they can become damaged and fuzzy. On the one hand, this may impede the interpretation of the photograph; on the other, it may give added authenticity or poignancy to the photograph.

In deciding which images to use, the researcher can ask participants to select images from their own or researcher-provided images, or the images may be selected on the basis of sampling techniques, for example, random stratified sampling of images, representative sampling, convenience sampling, probability and non-probability sampling from a given population, and so on. Strict sampling may not be possible if the still images are in very short supply (e.g. only one or two images are available). Nonetheless, as with other forms of data and participants, the selection of which images to use is subject to specification of criteria; the selection may be made on objective grounds (e.g. researcher-specified criteria or those which derive from the research questions), or subjectively from the participants themselves (e.g. their preferences or selections). The researcher/participant should specify and justify the selection made. Further, it is important for the researcher to elicit a narrative from participants, to explain the photograph (including, for example, factual and non-factual matters), as a single photograph may contain multiple messages (Elliot *et al*., 2016).

Researchers can consider several further questions in using photo-elicitation (cf. Barton, 2015; Richard and Lahman, 2015; Elliot *et al*., 2016), for example:

- what instructions to give to participants about taking and interpreting photographs, their focus and purposes;
- how 'provocative or disruptive' (Barton, 2015, p. 198) to make the photographs, when provided by the researcher;
- how to select the most suitable set of photographs: who decides and on what grounds;
- when to use the photographs (the most suitable time);
- how deep an impression the photographs may make on the participants;
- what difference it makes to the research and its outcomes if the researcher and/or participants take and choose the photographs;

- how to avoid doing harm to participants in photo-elicitation (e.g. evocation of intense negative memories);
- how important are the photographs in the research;
- what is the place of the photo-elicitation in the overall research and in large-scale (e.g. quantitative) data collection;
- how to combine a photograph with a narrative (the researcher's and/or the participants');
- who will see the results and in what form (e.g. publication).

31.4 Video and moving images

Taking and viewing moving images (film, video) are part of the everyday lives of everyday people, be they members of a family, the public, researchers, security and surveillance services or others. Video catches real-time sequences and behaviours in a clear chronology, often in close detail with high granularity, and can be stored and shared easily (e.g. with smartphones) (Blikstad-Balas, 2016). As with photographs, it may be the researcher or the participants who create the video, the latter being an instance of participatory research (Jewitt, 2012). Indeed the advantages of authenticity, ownership and empowerment claimed for having participants create photographs apply equally to participants creating videos, and having participants create the video can enable the researcher to gain access to their lifeworlds (Jewitt, 2012, p. 8).

Video has the attraction of recording 'naturally occurring' behaviour and events (Jewitt, 2012, p. 4), and, as with photographs, it has considerable evocative potential, re-awakening memories and events in participants, heightened by the multi-sensory, colour-rich moving image, i.e. creating or restoring a feeling of what it was like to be there. Further, the researcher can watch and re-watch the video, pause and freeze-frame, edit, remove or restore the sound, and focus in close-up detail on items.

Video material catches the non-verbal data that audio recordings cannot, which may be particularly useful, for example, in detailed case study data collection (e.g. of children at work, at play, interacting with each other and with adults). Video material is live and is useful for recording evolving situations and interactions, details that the observer may miss, and non-verbal matters (e.g. facial expressions, aggressive behaviour) (Greig and Taylor, 1999, pp. 66–7). It allows for repeated viewing and checking, though this takes time to watch, re-watch and analyse. The construction and consumption of video, as a meaning-laden resource, can promote 'reflective, dialogical and dialectical' thinking (Hadfield and Haw, 2012, p. 323).

Flick (2009, p. 249) reports the use of video for catching: (a) natural social situations; (b) contrived situations, for example, experimental conditions and situations, events and activities as recorded by the participants themselves and/or the researcher; (c) posed situations (such as video diaries); (d) special events; or (e) commissioned materials (e.g. a celebration or commemorative activity).

As with photographs, the researcher has to be aware of the selective bias inherent in moving images, i.e. the images recorded are a function of the focus and location of the camera, as well as the editing of the material. Hence the researcher must consider not only the images themselves and where, how, why, for whom, how and under what conditions they were produced, but also the interpretations that he or she (or indeed others) make or may make of the moving images, and how these interpretations are influenced by the interpreters' own backgrounds, values and purposes, i.e. the issue of reflexivity.

Moving images are powerful in many kinds of educational research, from experiments to ethnography. They can catch both the everyday routines and practices of participants and also special events. On the one hand, they are rich in detail, and on the other hand, this raises problems of how to analyse complex and detailed, often superfluous multimedia data, in ways that do justice to the different media (sound and vision) both separately and together.

Video data are rich but they are also selective, shaped by decisions of the video maker (Jewitt, 2012, p. 8), and this risks bias. The video material depends on the focus and angle of the camera, whether it is a fixed camera (the 'eye in the classroom'; see Chapter 26) or moved round the location and focused by an operator, a wide-angle lens or a lens with close-up focus, and indeed it depends on when and for how long the camera is taking the moving images. Whilst videos are rich in detail, this presents issues of how to conduct and write up the data analysis. Flick (2009, p. 250) also draws attention to important legal and ethical matters of permission, data protection, privacy, covert research (on the public and on identified persons) and permission to film (see the discussion below on the ethics of taking and using visual data).

A fixed camera in a classroom is not neutral; it has its field and focus predetermined. A wide-angle lens might catch gross behaviours but miss important detail – an eye movement, a facial expression, a small hand movement, a finger gesture. A fixed camera may be less intrusive, as it does not need the presence of an

operator and, indeed, may be located in a ceiling-level corner of the classroom. However, people move in and out of the field and focus of a fixed camera. Sometimes the video camera might be supplemented by a microphone situated on the table(s) at which children/participants are seated. It is important, therefore, in using a fixed camera, to decide where to locate it and focus it, whether there will be a camera operator or whether the camera will run automatically.

Having a moving camera that is operated by a person *in situ*, whilst it may catch close-up detail, may be highly intrusive and artificial (though Blikstad-Balas (2016) suggests that reactivity may be overstated, particularly if the camera is present for a longer period of time). In taking moving images, the researcher will need to consider the location, height, visibility and intrusiveness of the camera, the field of focus, the lighting in the area to be filmed, when to start and stop the recording (and whether the recording will be continuous or intermittent), how may cameras to use, whether to have a fixed or moving camera, who will operate it and who will create/edit the video (Jewitt, 2012).

Given their selectivity, researchers often use videos in conjunction with other kinds of data (triangulation). Indeed Flick (2009, p. 252) advocates the use of video material as part of a wider database and methods rather than being stand-alone. Data from moving images can be used for discussion (e.g. in subsequent interviews), to ask for video participants to reflect on the material, to corroborate data from other sources and to exemplify and illustrate themes, issues and events.

In using video, Blikstad-Balas (2016) notes the importance of addressing three challenges:

- balancing attention to close-up detail and the broader context, so that the broader context is not lost;
- avoiding data overload ('death by data') (p. 6), particularly magnifying events or details which might not be meaningful or important to participants;
- representing data, ensuring that audiences are able to judge if inferences that are made from the video are plausible (p. 5).

Because video data are complex, it is difficult to make sense of them and to represent them (p. 7), as video material is selective (and indeed can be edited by the video's creator) and therefore inevitably omits certain details or events. Even if one has more than one video camera trained on an event (e.g. inside a classroom), this does not overcome problems of interpretation and representation, as video data are inherently ambiguous and can sustain multiple interpretations. What the video frames is a selection only (p. 8), and attention to micro-matters, one of the advantages of video, might too easily overlook macro-settings. Many videos typically record short events and interactions (Jewitt, 2012, p. 8); they are not feature-length documentaries. Even though they tend to record short events, activities or behaviours, Jewitt (2012) notes that the risk of data overload is serious in video research, overwhelming the researcher, and this can weaken the research, rendering it descriptive rather than analytical (p. 6).

As with photographs, it is essential for the researcher (and indeed the participants) to accompany the video with the construction of a narrative that makes meaning of the video, as, without this, there is no evidence of intentionality, what is in the minds of the researcher/participants, opinions, significance, interpretations, motives of participants, i.e. all those points which a material, observable image cannot include.

Useful sources for using moving and still images in research can be found in: Heath and Hindmarsh (2002); Flick *et al*. (2004); Knoblauch *et* al. (2006); Banks (2007); Pink (2007); Rose (2007); Konecki (2009); Heath *et al*. (2010); Mitchell (2011); Jewitt (2012); and Blikstad-Balas (2016). For online guidelines on conducting video research more specifically in education, we refer readers to the companion website.

As with still images, in deciding which moving images to use, the criteria for selection (sampling criteria) should demonstrate fitness for purpose, fairness and defensibility. Moving images may focus on, for example, critical incidents, turning points, key events, representative behaviours, extreme examples, and so on. The criteria for choosing video clips must be justified. As with still images, the moving image clips may be selected by the researcher or the participants, and, as with still images, strict sampling may not be possible if the moving images are in very short supply (e.g. only one or two videos are available). The researcher should specify and justify the selection made.

The use of video raises serious ethical issues, as posting videos in the public domain – the sharing of the video – is very easy, raising issues of privacy, confidentiality, anonymity, protecting people from harm, ownership of the image, informed consent (and for what). We discuss these below.

31.5 Artefacts

As with other visual data, objects/artefacts can convey messages, even if those messages are unclear. Artefacts include, for example, objects in interior design and equipment (Higgins and McAllaster, 2004), children's toys, reading materials, DVDs, clothes etc. (which can

give indications of gender stereotyping in young children and how such stereotyping occurs and how boys and girls are inducted into differently gendered worlds (e.g. Francis, 2010)).

Artefacts have been widely used in educational research (e.g. Boston, 2008; Francis, 2010), ethnographic, anthropological and historical research, and studies of organizational culture (Schein, 1992), for example, studying dress codes, architecture, status symbols, signs, furniture, office areas, space, technology, mission statements and physical premises (Buch and Wetzel, 2001). Schein (1992) considers artefacts to be one of the three main levels and manifestations of organizational culture. Artefacts are the observable level of organizational culture (the other two levels being values and deep-seated norms); they are the outward manifestations of culture, for example executive rooms, dress codes, level of technology utilized (and where it is utilized), the physical layout of workspaces, the objects provided or observed in the workplace. All may be visible indicators of culture, but they are difficult to interpret; artefacts may suggest what a group is doing, but not why.

For example, consider a dull, dark, sparsely fitted classroom with no real amenities or decoration, with a few dried pot plants in a corner and no surplus ornaments or displays. Contrast this with the brightly lit, interesting, multi-equipped classroom with notices, displays, samples of students' work and the latest computing equipment in use. The objects can make a point here very tellingly, but what is that point? Is it that:

- some classrooms are dull, dispiriting places whilst others are energizing and interesting;
- some schools don't care about the teaching room whilst others take pains to present a stimulating environment;
- some schools are financially poor whilst others are rich;
- some classrooms exude a focus on learning from the teacher whilst others emphasize learning from the environment;
- some classrooms do not care about students' emotions whilst others are concerned to make the environment a happy place;
- some classrooms are very old and off-putting whilst others are new and engaging?

Inference from the artefacts alone may be dangerous as they may signify very different or discrepant realities; hence researchers should consider using artefacts alongside other data sources.

Take the example of the school in which the principal's office is private, separated from the main part of the school, large, beautifully carpeted, airy and spacious, with trophies, pictures, gifts, a huge working desk and an ergonomically designed chair, maybe a glass cabinet or two, works of art, a photograph of the family and of a meeting with an important dignitary, an up-to-date computer and colour printer, and a personal bathroom. Contrast this with the working space of the staff, who each have a small cubicle as part of a large room which has been sectioned off into workspaces for a dozen or more staff, with eye-level partitions like a typing pool, a small chair and desk, no room to put anything personal, with workstations squashed into an egg-crate arrangement and with no personal space, no superfluous ornaments, not a picture in sight, bare walls except for notices, shared equipment, and piles of books in each cubicle waiting to be marked. The messages – the not-so-hidden curriculum – of power, status, care and respect for people and humanity are very clear.

Or take the example of a school staffroom which may be untidy, with piles of books strewn around in different places, unwashed cups all over the room, notices peeling off the notice boards, cushions crumpled up on chairs, boxes of sports equipment lying in the corners, box files piled up alongside tables, comfortable chairs in very short supply and pieces of computer equipment cluttering up several tables. What can the researcher infer from this scene: that staff are extremely casual and careless or that they are extremely busy? Very different interpretations can be made of the same scene and artefacts.

As with other visual materials, artefacts can give messages but, like other visual objects, they are easy to observe but difficult to interpret, and there are multiple interpretations. In some cases artefacts may be easy to interpret, for example, the images presented in children's books may indicate sex role stereotyping, or may portray positive images of some groups and negative images of others.

Artefacts can be seen, heard, smelt, touched, felt, even tasted and heard, so the researcher can bring to bear a multi-sensory analysis. They can be used to stimulate discussion (see the comments above about the use of photographs), to glimpse into the past or the present, to reconstruct or help to imagine a scene, to remind people and bring back memories. They can be observed *in situ* (and the location and placing of the object in a spatial context itself carries meaning, e.g. in a home, a museum, at the back of a room, in a dark corner, in a prominent position etc.). As with still and moving images, the artefact may be provided by the researcher or by the participants.

The researcher can examine artefacts on their own or in combination. For instance, in the example of the

messages about the organizational culture of the principal's and staff's office areas, a single object may not say very much, but taken together the objects can make a persuasive case. Objects may also be sorted and grouped into categories, for example, ornaments, books, furniture, space; each category can be examined on its own and/or in combination (akin to the different kinds of coding exercise in grounded theory, where individual codes are combined into categories) and such sorting can elicit the participants' conceptual categories: how they think and group items, and why (cf. the earlier discussion of elicitation) (Barton, 2015).

In investigating artefacts, the researcher can consider the purpose of the production and location of the artefact, what it was used for and by whom, who produced it, when was it made, what materials were used in its making, what was its actual and/or symbolic purpose or function, how has it been preserved and in what condition, and what value it has to the provider or user. This has particular significance in historical, anthropological, ethnographic and archaeological research in educational and social science.

In some research (e.g. on child abuse), artefacts (e.g. dolls with lifelike features or sex organs) can be used to encourage children to speak out about their experiences, displacing the highly sensitive personal threat or embarrassment onto the doll in question. Greig and Taylor (1999, p. 64) advocate the use of familiar artefacts with children – dolls, puppets, drawings, pictures – as this not only sets them at their ease but helps them to make concrete their ideas. This technique is particularly useful with young children, where dolls or puppets can have a series of facial expressions (happy, sad, angry, afraid) and where non-verbal postures can be manipulated on puppets (e.g. dolls, manikins, glove puppets) to enable the researcher to investigate emotions in young children (pp. 120–2). Greig and Taylor indicate how puppets can be used to research situations of conflict in young children. For example, the researcher can ask what puppets A and B want, how they feel, why they are fighting, who is winning, whether the fight is justified, what each puppet should do, what the child would do in a similar situation, why puppet A or B was wrong, how the situation could end, and how the situation could be resolved (p. 122).

How researchers use artefacts depends on their research questions. Similarly, just as one uses sampling procedures to decide, for example, which people to approach to be involved in the research, so one has to consider the criteria to be used for deciding the sampling and selection of artefacts (cf. Lodico *et al*., 2010, p. 164). As with still and moving images, in deciding which artefacts to use, the criteria for selection (sampling criteria) should demonstrate fitness for purpose, fairness and defensibility. This operates with researcher- or participant-provided artefacts and with researcher observation of artefacts (e.g. the objects in a classroom, staffroom, principal's office and so on). The researcher should specify and justify the selection of the artefacts, and, to be faithful to the multiple interpretations that can be made of artefacts, the researcher should provide alternative interpretations of the artefacts where appropriate.

31.6 Ethical practices in visual research

Taking visual images is subject to the same ethical concerns and requirements as other forms of educational research, and we refer readers to Chapter 7 here. In particular, the issue of informed consent may prove difficult in the case of historical images, images of the general public or deliberately covert research. It is important to consider the indiscriminate taking or use of photographic or visual images of children without their consent and that of their teacher, the school, parents, helpers, guardians and staff.

Whilst Clark (2006) and Prosser *et al*. (2008) regard collaborative research (between researcher and participants) as one way of addressing complex ethical issues, this does not cover all situations, and researchers need to consider the ethical principles set out in Chapter 7.

Public places

Permission concerns not only the site of the image itself (e.g. the taking of the photograph, filming in public places), but permission for reproduction (e.g. from individuals, from institutions), indicating the uses to which the image will be put, and indeed for altering the image in some way. In the case of public places (and Prosser *et al*. (2008, p. 6) argue that what constitutes a public place is itself unclear), permission may need to be sought from the official body or party responsible for that public place as well as individuals (e.g. the informed consent of people in the street or in a building). Not only are there issues of legally and illegally taking images (e.g. of military establishments) or storing images, there is the issue of preferred and non-preferred sites for taking pictures (Prosser *et al*., 2008, p. 6), such as police stations, hospitals, schools, leisure facilities, surgeries, even rail stations, airports and libraries.

However, the argument is not only about what is permitted or not permitted in terms of a public place, and what constitutes a public place, but the ethical acceptability of, for instance, taking a photograph of

someone in a public place without their informed consent. What constitutes a public, semi-public or private space is unclear, or what constitutes what the participants may wish to see as a private activity in a public place (e.g. kissing your partner) (cf. Solove, 2004). For example, just because there are no notices in a hospital or a school to saying 'no photographs', does that make it ethical to take such photographs or videos of patients, the public, students and teachers in those sites (cf. Clark *et al*., 2010)?

Anonymity, confidentiality, privacy and identification

Images, for example, photographs and videos, often identify people. This raises significant ethical issues. Identification, anonymization and obscuring of individuals and places relate not only to the ethical sphere but to matters of legal regulation on data protection. On the other hand, some participants may deliberately wish to be identified (Prosser *et al*., 2008, p. 11; Brooks *et al*., 2014, p. 145).

Prosser *et al*. (2008) contend that visual methods raise issues of informed consent, anonymity, confidentiality and dissemination (p. 2). Anonymity and confidentiality may be highly problematic in visual images, as the whole purpose of the image lies in the details of the person, place or institution in question, without disguise or dehumanization (p. 15). If one removes identifying features, one destroys the very detail that might be the purpose of the research.

Whilst informed consent may be one method of addressing this, it is not always applicable (discussed below). Images may be anonymized, for example blurring identifying features, pixilation, eye-blocking (e.g. black bars), voice modification, using pseudonyms, taking the image showing only the back of the person, showing only non-identifying features of people (e.g. their hands), or with shaded, back-lit lighting (see Clark (2006) and Clark *et al*. (2010) for detailed guidance on anonymization).

Copyright and ownership of the image

Visual images may be subject to copyright and intellectual property legislation. Researchers have to be aware of their moral and legal responsibilities concerning intellectual property. For example, if one takes a visual image of a teacher and students in a classroom, who owns the image and the lesson in question (US Department of Education, 2002) – the researcher, the teacher, the students, the parents, the school, or any combination of these – and what rights to usage, distribution and publication does this bring or prevent? What if the image includes a school textbook: does the publisher have ownership rights? What if the students have special needs? What if the class is unruly? What if the lesson is ineffective? How can the data be used, and by whom? Once the image is in the public domain, it is beyond the control of the researcher.

'Images' come under legislation concerning 'artistic works' (Clark *et al*., 2010), with copyright lasting for decades after the artist's death, and the researcher's use of photographs is affected by Data Protection Acts which require informed consent, not least for public release. Informed consent applies not only to taking the photograph or video, but to putting the image into the public domain. This can be particularly problematic, for example, if children or researchers have created images or collages of images using media-created photographs (e.g. from magazines).

Informed consent

Informed consent (see Chapter 7) is complex, as, for example, it may be in the public interest to have a long-lens photograph of a private activity or of a person, without that person's informed consent, i.e. covert research. As discussed in Chapter 7, informed consent also raises issues of who is to give that consent, and for what, and on whose behalf. Clark *et al*. (2010) note that this is a delicate matter, as seeking consent from parents may disempower their children who are the very subjects of the research. Further, practically speaking, it may be impossible to obtain the consent of everyone who appears in the image, for example, a crowded school corridor or playground. Informed consent applies not only to the creation of the image (e.g. the photograph, the video) but to its use, release, publication and audiences. The long-term effect of publication is relevant here: for example, Clark *et al*. (2010) report the case of a researcher who decided not to publish a photograph, because in later years the topic could come back to haunt or negatively affect the person in the photograph. Once an image goes into the public domain, the researcher almost completely ceases to have control over how it will be used.

Prosser *et al*. (2008) give an example of the challenges that visual researchers face with regard to informed consent (pp. 12–14):

- it may not always be appropriate to gain informed consent (e.g. in covert research or surveillance work);
- what 'informed' and 'consent' mean may be different in different cultures or with different groups (e.g. children);
- it is not always clear who is actually in a position to give the consent sought (e.g. in the case of children or teachers);

- it may not be practically possible to gain the consent of those who feature in visual images (e.g. in public places), for instance in the case of photo-journalism;
- it may be difficult to gain the consent of those featured in a visual image if the provider of the image (e.g. a participant) has not gained that consent;
- it is important to ensure that participants know to what they are giving their consent, for example, to the taking of the image, to the reproduction of that image (and where) and for how long;
- it is not always clear what to do with 'found images', where the provenance of the image is unknown or with images which were not originally produced for the purposes used in the research (see the analysis of the photograph in Chapter 36).

In covert research (see Chapter 7), with easy access to photographing and videoing with smartphones, or a range of small video cameras, disguised and tiny, issues of informed consent are highly problematical.

Do no harm

A key principle in image-based research is *primum non nocere:* 'first of all, do no harm'. This issue concerns the sharing and publication of the image, and with whom (Alderson and Morrow, 2011; Hammersley and Traianou, 2012), and researchers have to consider not only legal constraints but ethical obligations. They must ask themselves, and indeed the participants in the research, what might be the negative effects of the image on those included in the image, on the participants and on the researcher, if the image is made public, both now and in the future. What if there is an image of students being unruly or if the teacher is ineffective? What if the image, for example, a photograph or video, is taken covertly? What if a person does not wish to be photographed or videoed or doesn't know that this is happening? How can the data be used, and by whom? Once the image is in the public domain, it is beyond the control of the researcher, and *primum non nocere* trumps benefits to the researcher; this is an overriding precept.

A statement on ethical practice in visual research is provided by the British Sociological Association (2006). This includes: professional integrity; legal considerations (including data protection, copyright and libel laws); ownership of images; images of illegal activities; morally questionable practices; beneficence and non-maleficence; non-breaching of trust; informed consent; relations with and responsibilities towards research participants; sensitivity to local cultures; procedures for sharing images; covert research; researching vulnerable groups; anonymity, privacy and confidentiality; dangers of intrusion into private worlds and lives; working with children and images of children; Internet-based research; relations with and responsibilities towards sponsors and/or funders; and clarification of rights to publish.

The UK's Economic and Social Research Council National Centre for Research Methods (2008) has produced a comprehensive analysis of ethical issues in visual research, including material on frameworks, professional guidance, regulations and legal rights and duties for visual researchers. It covers: ethics; issues of consent; researcher-generated and respondent-generated images; anonymizing and obscuring visual data; photo-elicitation and informed consent; anonymity and confidentiality; photographs and films that identify individuals; images of place and how to anonymize these; the construction and consumption of images; and guidelines for practice. We strongly advise researchers to consider carefully these ethical guidelines, as they indicate the very careful boundaries within which researchers with visual data must work. We provide the websites of these organizations on the companion website, and we also refer readers to Clark (2006), Wiles *et al*. (2008), Prosser *et al*. (2008) and Skåreus (2009).

We summarize key questions in image-based research and visual methods in Boxes 31.1 to 31.4. 'Image' here can refer to photographs, videos, drawings, artefacts etc., i.e. the entire gamut of visual objects.

BOX 31.1 APPROACHING IMAGE-BASED RESEARCH

In the research design

1. How does the use of the image fit with the purpose, focus and design of the research?
2. What research question(s) will the use of images address?
3. Why include image-based research?
4. What are the purposes and importance of creating and using the image in the research?

Preparing the image and its usage

5. What visual medium will you use? What kind of image is it (photograph, video, collage, painting, drawing, graphic etc.) and why choose that kind?
6. Is it an existing or specially created image?
7. How many images?
8. How and why is/was the image created/selected/included?
9. Who creates/created the image?
10. Who provides the image?
11. What is the context of the image?
12. Why and how was the image originally created and obtained?
13. What equipment and training are needed to create the image?
14. What is the source of the image (e.g. book, magazine, archive)?
15. What kind of image is it (e.g. a mirror/real event, a metaphor or symbol, a representation, a narrative sequence, concrete or abstract etc.)? What is the image about?
16. What instructions, preparations and criteria have been given to participants in connection with the image (e.g. creation, usage, selection)?
17. What are possible effects of creating the image on participants and those included in the image? How to address intrusion and reactivity?
18. How to address researcher and/or participant reflexivity in creating and using the image?
19. How to avoid researcher and/or participant bias in creating and using the image?
20. What technologies, compositional and editorial features have been used to create the image, and what effect does this have on the image?

BOX 31.2 USING THE IMAGE IN THE INTERVIEW

Using the image in the interview

1. When and how to introduce the image in the interview, and why?
2. How to use the image to build rapport and trust?
3. How to use the image to promote discussion/data collection (i.e. what to say, how to prompt and probe)?
4. How to respect the 'positionality' of the participants in interpreting the image?
5. What narrative/commentary are you seeking in the use of the image?
6. How to respond to participants' reactions (e.g. emotional) to the image and its use in the interview?
7. How to balance attention to close-up detail with attention to the bigger picture?
8. How to ensure that the points raised about the image at interview are relevant to the participant?

BOX 31.3 DATA ANALYSIS WITH IMAGE-BASED RESEARCH

After the interview

1. How to analyse the images and the interview data?
2. What to focus on the analysis, and why?
3. How to create and include narratives about the image?
4. How to report discussion/interview data, to whom, in what form, and with what protections?
5. How to incorporate image-based analysis with analysis of other data?
6. How to address researcher and/or participant reflexivity in reporting the results?
7. How to avoid researcher and/or participant bias in reporting the results?
8. How to avoid selection bias in what is reported?
9. What are the main messages from the interview/image analysis, and how do we trust these?
10. How to check and make transparent the plausibility of the researcher's interpretations of the image and the interview?

BOX 31.4 ETHICS AND OWNERSHIP OF IMAGES

Ethics and ownership of images

1. How has the ethic of 'do no harm' been addressed in creating and using the image?
2. What informed consent has been obtained, from whom and for what?
3. How have issues of anonymity, confidentiality, privacy, non-traceability and identifiability of people, places, institutions, behaviours and events been addressed?
4. Who will see the image?
5. How will the image-based research empower participants?
6. How to use images in sensitive research?
7. What safety precautions have been taken in creating, using, storing, sharing, disclosing and publishing the image (legal and ethical)?
8. Who owns the image?
9. Where to post images on the Internet/cloud and with what security features?
10. What permissions have been obtained for access to image-creation sites and control of release of the image (including legal requirements)?
11. What copyright issues have been addressed?

Companion Website

The companion website to the book includes PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. These resources can be found online at **www.routledge.com/cw/cohen**.

Part 5
Data analysis and reporting

This part starts with qualitative data analysis and then moves to quantitative data analysis. In qualitative data analysis we take readers from first principles to content analysis and grounded theory, making the point that texts – data – are multi-layered and open to a variety of interpretations. We indicate in practical terms how researchers can analyse, present and report qualitative data, including an introduction to the foundational principles of such approaches and ways of organizing and presenting qualitative data. The material on qualitative data analysis covers kinds of analysis which are not reliant on coding and grounded theory, including conversational analysis, narrative analysis, discourse analysis and autobiographies. We provide many worked examples of each of these, indicating how to approach, conduct and report analyses of different kinds of qualitative data, and the need for reflexive authorship of analyses. We expand the chapter on analysing visual data, including new examples of software for approaching and analysing text, still images and moving images.

In quantitative data we assume that researchers will not only have no experience of statistics but may even be frightened by them! Hence we take readers by the hand from very first principles to more complex statistical processes, how to do them and how to analyse, interpret and report the results. We explain the foundations, principles and concepts underlying statistical procedures, and we indicate 'safety checks' that researchers should follow before proceeding with statistical analysis. We have an entirely new chapter on statistical significance, effect size and statistical power, how researchers can work with these, and some concerns about significance testing. In introducing descriptive statistics we also include new material on missing data.

Following this we introduce correlations, reliability, straightforward inferential statistics, including difference tests, regression and working with standardized scores, and higher level inferential statistics, such as factor analysis. We include new material on cluster analysis and introductory notes on, and examples of, structural equation modelling and multilevel modelling. Given the range of statistics available to researchers, we organize these clearly using charts and tables, so that researchers can see how to select appropriate statistics for their purposes and to fit the kinds of data collected. Finally, this part includes an entirely new chapter entitled 'Beyond mixed methods: using Qualitative Comparative Analysis (QCA) to integrate cross-case and within-case analyses', which is a fitting final chapter as it works with a celebrated approach to moving beyond discrete quantitative and qualitative data analysis. These chapters also contain new boxes indicating the command sequences for using SPSS.

For both the qualitative and quantitative data analysis we provide practical advice – including sample phrases and choice of words – on how to report results and findings. These chapters are accompanied by extensive materials on the companion website, including data files for use with qualitative and quantitative data and PowerPoint slides for each chapter, at: www.routledge.com/cw/cohen.

Approaches to qualitative data analysis

CHAPTER 32

This chapter sets the scene for qualitative data analysis. Subsequent chapters focus on specific techniques and types of data analysis such as coding and content analysis, discourse analysis, narrative analysis and grounded theory. Here we introduce:

- elements of qualitative data analysis
- qualitative data analysis, thick description and reflexivity
- ethics in qualitative data analysis
- computer assisted qualitative data analysis (CAQDAS)

In many instances we deliberately use seminal texts in this field. In-depth consideration of ways of analysing qualitative data is set out in Chapters 33 to 37.

32.1 Elements of qualitative data analysis

Qualitative data analysis concerns how we move from the data to understanding, explaining and interpreting the phenomena in question (Taylor and Gibbs, 2010, p. 1). It includes, among other matters, organizing, describing, understanding, accounting for, and explaining data, making sense of data in terms of the participants' definitions of the situation (of which the researcher is one), noting patterns, themes, categories and regularities, all of which are the task of the qualitative.

Qualitative data analysis is not straightforward. As Patton (2002) remarks, though qualitative data analysis turns data into findings, there is no simple formula or recipe for this (p. 432). There is no one single or correct way to analyse and present qualitative data; how one does it should abide by *fitness for purpose*. Qualitative data analysis is often heavy on interpretation, and there are often multiple interpretations to be made of qualitative data – that is their glory and their headache! It is also distinguished by its merging of analysis and interpretation and often by the simultaneity of data collection with data analysis (Gibbs, 2007, p. 3) in a back-and-forth process (Teddlie and Tashakkori, 2009, p. 251). Indeed, the processes of the analysis also constitute data in themselves. As researchers write down notes, memos, thoughts and reflections in the field or during an interview or observation, these, too, become data. The steps taken in moving from description to understanding to explanation, interpretation and conclusions must be transparent (Gläser and Laudel, 2013) and demonstrate validity.

Qualitative data derive from many sources, for example:

- interviews (transcribed or not transcribed);
- observation (participant to non-participant);
- field notes;
- documents, reports, newspapers, minutes of meetings;
- memos;
- emails and online conversations;
- diaries;
- audio, image-based, visual, video and film materials;
- records of events;
- websites and website data (e.g. online surveys);
- qualitative survey data (e.g. from questionnaires);
- advertisements and print materials;
- pictures and photographs;
- artefacts.

The list is huge. Qualitative data analysis focuses on in-depth, context-specific, rich, subjective data and meanings by the participants in the situation, with the researcher herself/himself as a principal research instrument. It involves: *data reduction* (in order to avoid the often serious issue of data overload, i.e. too much detail and too much material); *data display*; *data analysis and interpretation*; *drawing and verifying conclusions*; and *reporting* the analysis and findings (cf. Miles and Huberman (1984, 1994, who restrict their suggestions to: data reduction, data display and conclusion drawing and verification). Data reduction does not mean disregarding data; rather it means distilling from the complexity of the findings the key points of the phenomenon in question, reducing complexity without violating it, catching the essence of the issue or the situation, enabling the researcher to identify, for example, patterns, key issues, causal processes and sequences (Gläser and Laudel, 2013).

To compound the challenge of qualitative data analysis, as Chapter 15 made clear, meanings and interpretations of situations and data are not singular or unitary – the sole preserve of the researcher(s) – but are several, including those of the participants. In other words, there are many possible analyses and interpretations of data, and care must be taken to avoid indefensibly privileging one interpretation over another equally possible interpretation (e.g. the researcher's) if both are sustainable by the data. The researcher has to catch multiple perspectives on a phenomenon and multiple interpretations of it, and report these.

Qualitative data analysis involves some or all of the following, depending on the approach adopted (e.g. Newby, 2010; Creswell, 2012; Gibbs, 2012; Marshall and Rossman, 2016):

- *preparing and organizing the data*: putting the data into formats, documents, maybe computer files, and an organization system that make for ease of management and analysis. This might involve transcribing and/or summarizing data, and bringing order into the data;
- *describing and presenting the data*;
- *analysing the data*: exploring and making meaning of the data, for example, organizing and categorizing data into key concepts; identifying the units of analysis; coding; inductive processes; identifying and refining key concepts and key points; identifying linkages and relationships between the data; summarizing; thematic analysis; creating typologies (e.g. by combining categories – the qualitative equivalent of factor analysis from variables in quantitative research); case summaries and cross-site analysis; patterning; constant comparison methods; discourse analysis; writing a narrative, conversational analysis etc. In other words, it involves data assembly and re-assembly, recombining them in new ways, synthesizing and integrating data in order to create a meaningful account and analysis;
- *interpreting the data*;
- *drawing conclusions;*
- *reporting the findings*;
- *ensuring accuracy, reliability, coherence, corroboration, validity and reliability* (variously defined, see Chapter 14).

These bullet points overlap and do not necessarily indicate a sequence, indeed analysis and interpretation often occur simultaneously. Rather, the process of data analysis is recursive, non-linear, messy and reflexive, moving backwards and forwards between data, analysis and interpretation. It involves ensuring that all the relevant data have been included and that a fair, coherent and defensible representation of the data and their meaning(s) has been presented, together with conclusions drawn from them.

Data analysis is often an ongoing process that takes place during the research as well as at the end of it. Data collection and analysis may accompany each other (see below: progressive focusing), and this means that analysis is subject to continual modification, addition, refinement, excision, extension and amendment. Some of the analytical tools can be *pre-ordinate* (*a priori*: ideas, themes, codes, key points, analytical framework etc. decided in advance); some can be *responsive* to the emerging data and their analysis and interpretation (*a posteriori*) and what they reveal; indeed a combination of *pre-ordinate* and *responsive* categories, codes, themes, ideas, topics, concepts etc. can be used.

The qualitative data analyst must identify and locate raw data (Gläser and Laudel, 2013, p. 5) and link data to the research questions and the research findings (Thomas, 2006). This involves interpretation, and categories generated can be derived from the data themselves, or from theory, research questions, the researcher herself/himself or any combination of these. Similarly, linkages between data can be in terms of concepts, themes and content, with patterning across data (i.e. more-than-one occurrence of sequences, conditions, processes and outcomes of events), or indeed in terms of conflicting accounts. Such patterns can be integrated or merged where suitable (Gläser and Laudel, 2013, p. 8), ensuring that all the data are included and none forced out, even if this means acknowledging variation within an overall pattern.

Wellington (2015, p. 263), suggests that qualitative data analysis includes: (a) dividing the data into 'units of meaning'; (b) classifying and grouping the units of meaning; (c) including new units of data into these groupings/categories; (d) searching for categories that are similar and/or which can be merged into a single category; (e) reviewing categories that contain large amounts of data to see if they can be split into smaller categories; (f) checking that the categories include all the data and are mutually exclusive (though some data may appear in more than one category); and (g) looking for linkages, contrasts and comparisons between the categories (constant comparison).

He provides a seven-stage model for 'making sense of qualitative data' (2015, p. 267):

Stage 1: 'Immersion' in the data;
Stage 2: 'Reflecting, standing back';
Stage 3: 'Analysing' ('dividing up, taking apart, selecting and filtering, classifying, categorizing')

Stage 4: 'Synthesizing, re-combining' data;
Stage 5: 'Relating to other work, locating' data;
Stage 6: 'Reflecting back (returning for more detail?)';
Stage 7: 'Presenting, disseminating, sharing' the findings.

This sequence is not necessarily linear, and recursion might occur.

In preparing for data analysis, the researcher must immerse herself/himself in the data, read, re-read, reflect on the data, write about the data and what they mean (and what different meanings, explanation and interpretations of the data there may be), how the data are linked or related, how to organize the data and the key points arising from the data, how to analyse the data, how to organize and synthesize the analysis most fittingly and coherently (e.g. Wellington, 2015), and reflect on how the researchers' own biography, values, knowledge, assumptions and experiences shape or inform the data analysis, i.e. reflexivity (Woods *et al*., 2016, p. 387).

Qualitative data analysis as an inductive process

The process of qualitative data analysis is typically inductive (Thomas, 2006). Here the researcher reads, re-reads, reflects on, infers from and interprets the raw data/transcripts/memos etc. From this, without preconceptions or deductions from a pre-given framework (unlike, for example, experimental research or hypothesis testing), the researcher develops interpretations of the data and derives themes, concepts, theories, explanations, understandings, summaries, models etc. which fairly and comprehensively explain the data or phenomenon. As Strauss and Corbin (1998) remark, the theory emerges from the data in the field of study in question (p. 12). This is a bottom-up process, moving from data to explanation to theory (hence the 'grounded theory' approach discussed in Chapter 37) (though Bazeley and Jackson (2013) contest how far it is only a bottom-up process). The researcher can refer to the research questions in guiding induction, though often analysis is data-driven rather than research-question-driven (Thomas, 2006) (but see Chapter 33 here). The inductive process, Thomas suggests, can proceed thus (in practice this is not necessarily a linear sequence): *understanding the research objectives* → *preparation of the raw data* → *reading and re-reading the raw data*, and *reflecting* on the raw data and their meanings → *category generation, revision and refinement* (involving, as appropriate: coding and recoding; creating a hierarchy of codes and categories; checking for consistent use of codes; category descriptions; removing irrelevant data without loss of fidelity to the phenomenon; category labelling; identifying texts/data to put within the category; and identifying links between categories) → *model generation* (which might contain key themes, processes and the category systematization, temporal sequences and causal networks) → *answering the research purposes and questions*.

Many researchers move almost spontaneously into coding. We advocate caution here (see below and Chapter 34). Thomas (2006) suggests that checking for the consistent use of codes can be addressed by:

- *independent, blind parallel coding*, where a second researcher is given the objectives and the raw data, but no codes, and is asked to code the data, after which the two sets are reviewed for consistency, overlaps and discrepancies;
- *checking on the clarity of the categories*, where a second researcher is given the categories and the raw (uncoded) data and is asked to assign the raw data to the categories, which are then checked against the allocation of the data by the first researcher, to look for consistency and discrepancies.
- *stakeholder and member checks*: respondent validation and review in order to establish, *inter alia*, the credibility, transferability, dependability and confirmability of the data and the findings.

Thomas is writing about evaluative research here, and he emphasizes the need to understand the research objectives. In other research, for example, goal-free evaluation or open-ended research, the objectives may be less certain and less the engines of the research. He also recognizes that there can be more than one valid interpretation of the data and that interpretations are influenced by each researcher's own biography, experiences, values and assumptions; hence there may be more than one set of findings and conclusions, which may or may not be similar to each other.

Preparing and organizing the data: transcription and summary

Preparing the data means putting them into a format that lends itself to analysis. For example, it may be a matter of creating: (a) word files/text files of observational data, interview data, questionnaire data, memos, field notes and suchlike; (b) visual data files (e.g. of pictures, graphics, images, videos); (c) audio files; and (d) files of graphs and numerical data. All of these are prepared with ease of access, overall organization and data analysis in mind. This might mean putting files into a common format (layout, font size, organizational sequence, chronology, key features etc.), maybe with

computer analysis in mind so that they can be read into particular computer software. Many computer packages enable multiple kinds of files and data to be imported, and we discuss this below.

In preparing data files, researchers should consider whether to transcribe interview data for analysis. On the one hand, transcriptions can provide important detail and an accurate verbatim record of the interview. On the other hand they omit non-verbal aspects, what may take place before or after the interview and contextual features of the interview.

On a practical level, it may be difficult to catch exactly what was said if people do not speak clearly or sufficiently loudly, or they may speak in broken sentences or with different accents, and these features, whilst difficult for the researcher, may not appear in the final transcript (Denscombe, 2014, p. 279). Transcripts are also very time-consuming to prepare (e.g. one hour of interview may take up to five or six hours to transcribe, even with a transcription machine which can be paused whilst the transcriber writes the words or enters data into computer software). The researcher should consider the costs and benefits of transcription, judging whether close transcription is really necessary.

An alternative to transcription is to write the summary of data or their analysis directly from the video or audio recording, selecting out the important materials directly from the original source, thereby avoiding becoming so caught up in detail that sight of the bigger picture is lost. Such selectivity might also include verbatim quotations and short extracts, accompanied by the researcher's own annotations (e.g. informal observations and comments) (cf. Denscombe, 2014, p. 278). Nonetheless this is data interpretation and selection, and, as such, can reflect on the researcher as much as on the data.

If transcription is used, then the researcher must make clear the transcription conventions being followed (see also Chapter 35), for example:

- give each speaker a name or pseudonym (and keep a list separately of which speaker has which pseudonym);
- record hesitations, small to long pauses, and silences (e.g. through dots (…) in the text;
- record inflections and tone of voice (rising to falling), for example, write down the mood of the speaker or the speech at the time: anger, anxiety, sadness, excitement, questioning, hesitance etc.;
- note: the volume of the speaker (quiet, loud, whispering, shouting); the speed of the speech (slow, fast, hurried, calm); pitch (high, low); tone (calm, angry, excited, nervous etc.); breaks in speech (sudden, considered); stresses and phases in the speech; audible breathing out or breathing in;
- record non-verbal activity (e.g. standing up, leaning back) if transcribing from video recording;
- record uninterpretable noise (e.g. by using the words 'noise' or 'unclear noise' in brackets);
- record several speakers who are all speaking at the same time (e.g. with the word 'together' after each speaker's name or a uniting large bracket);
- be consistent in spelling (so that search and retrieval can be facilitated, particularly if software for this is used, discussed later);
- ensure that each line or section/paragraph is numbered (in Word this can be done through the 'Page Layout' menu);
- ensure that wide margins and double spacing are used for annotating text in hard copy form.

For a fuller description of these see Atkinson and Heritage (1999), Flick (2009, pp. 300–2) and Woods (2010). The transcriber must check the accuracy of the transcription, as it is not uncommon for speech to be heard incorrectly or for words to be confused Gibbs (2007, p. 19 gives many examples of such confusions).

Voice recognition software is available that recognizes and transcribes speech, and this can save time, though the reliability of the transcription is influenced by the accuracy of the speech recognition and the clarity of the speaker.

Managing data files and types

Data management and organization are important, maybe creating an index, for easy categorization and retrieval. Data can be of various types, for example, text, images, videos, graphics, audio, numbers, Internet sites and so on. Within each of these there are several sub-groups, for example, within text-based data there are field notes, interviews, observations, memos, documentary records, working notes and so on. The management of data is essential for ease of access, chronology, data types, linkages and retrieval etc.

The researcher should create a clear and easily accessible system for data location, storage, organizing, filing and handling, be it in terms of hard copy or soft copy. In effect, the researcher creates a mini-reference-library system, with different subjects in different library and shelf locations, different kinds of materials in different locations and indexing, all with secure storage and protected access. Or, put it another way, imagine a textbook, which has a table of contents, parts/sections, chapters, sections within chapters, and an index; the researcher has to create an analogical data storage, organization and retrieval system.

Software is useful here. For example, NVivo enables different kinds of *data* to be organized and stored by category (folders, sources), for example, interviews, observations, documents, videos, pictures, memos, images, audio, graphics etc., and, within each folder, there are specific data files. The software can also retrieve and store *searches* made of the data (e.g. by queries, by code, by node, by reports) and it can save files into different locations which contain *analysis* (e.g. text-based, graphics based); see Figure 32.1 for an example of this.

32.2 Data analysis, thick description and reflexivity

In abiding by the principle of *fitness for purpose*, the researcher must be clear what she/he wants the data analysis to do as this will determine the kind of analysis that is undertaken. The researcher can set out, for example:

- to describe;
- to portray;
- to identify;
- to summarize;
- to interpret;
- to discover patterns;
- to generate themes;
- to understand individuals and idiographic features;
- to understand groups and nomothetic features (e.g. frequencies, norms, patterns, 'laws');
- to raise issues;
- to prove or demonstrate;
- to explain;
- to seek causality;
- to explore;
- to test;
- to discover commonalities, differences and similarities;
- to examine the application and operation of the same issues in different contexts.

The significance of deciding the purpose is that it determines the kind of analysis performed on the data. This, in turn, influences how the analysis is written up. The data analysis is also influenced by the kind of qualitative study being undertaken. For example, a biography and a case study may be written as descriptive narrative, often chronologically, with issues raised throughout (e.g. Hamilton and Corbett-Whittier, 2013). An ethnography may be written as a narrative or stories, with issues raised, but not necessarily conforming to a chronology of events, and including description, analysis, interpretation and explanation of the key features of a group or culture (e.g. Mills and Morton, 2013). A grounded theory and content analysis can proceed through a systematic series of analyses, including coding and categorization, until theory emerges that explains the phenomena being studied or which can be used for predictive purposes (Glaser and Strauss, 1967).

The analysis will also be influenced by the number of data sets and people from whom data have been collected. Qualitative data often focus on smaller numbers of people than quantitative data, yet the data tend to be detailed and rich: thick descriptions (Geertz, 1973). Researchers will need to decide, for example, whether to present data individual by individual and then, if desired, to amalgamate key issues emerging across the individuals, or whether to proceed by working within a largely predetermined analytical frame of issues that crosses the individuals concerned. Some qualitative studies deliberately focus on individuals and the responses of significant players in a particular scenario, often quoting verbatim responses in the final account; other studies are content to summarize issues without necessarily identifying exactly from whom the specific data were derived. Chapter 33 discusses methods to be used with respect to people and issues.

Some studies include a lot of verbatim conversations; others use fewer verbatim data. Some researchers feel that it is important to keep the flavour of the original data, so they report direct phrases and sentences, as they are often more illuminative (diamonds!) and direct than the researchers' own words, and because researchers feel that they should be faithful to the exact words used. Indeed direct conversations can be immensely rich in data and detail. Ball (1990, 1994a) and Bowe *et al*. (1992) use a lot of verbatim data, not least because they interviewed powerful people and justice was done to the exact words that they used. By contrast, Walford (2001, p. 92), commenting on the 'fetish of transcription', admits that he 'rarely fully transcribed more than a few interviews for any of [his] research studies', not least because of the time that it took for transcription (he suggested a ratio of 5:1 – five hours to transcribe one hour of interviews – though it can take much longer than this).

At a theoretical level, a major feature of qualitative research is that analysis often begins early on in the data-collection process so that theory generation can be undertaken (LeCompte and Preissle, 1993, p. 238). Here researchers painstakingly take apart their field notes, matching, contrasting, aggregating, comparing and ordering notes made, then they set out the main outlines of the phenomena under investigation (pp. 237–53), then they assemble blocks or groups of data, putting them together to make a coherent whole (e.g. through writing summaries of what has been

found). The intention is to move from description to explanation to theory generation.

At a practical level, qualitative research rapidly amasses huge amounts of data, and early analysis can reduce the problem of data overload by selecting out significant features for future focus. Miles and Huberman (1984) advise researchers to start writing and analysing early and frequently (i.e. as soon as the first data have been collected, even in a longitudinal study), rather than leaving all the writing and analysis until the data collection or the study is over, as this enables 'progressive focusing', with selection of the key issues identified for further investigation. As Gibbs (2007, p. 25) remarks, 'writing is thinking'. Such analysis should, itself, be given a date and time, and can be included in a diary of field notes which record, for example, what the researcher was doing, where the researcher was, what was happening at the time, who was present, what the data were, particular or notable features of the event, context or situation, reflections and observation (Miles and Huberman, 1994, pp. 50–4).

'Progressive focussing' (Parlett and Hamilton, 1976) starts with the researcher taking a wide-angle lens to gather data. Then, by sifting, sorting, reviewing and reflecting on the data, the salient features of the situation emerge. These are then used as the agenda for subsequent focusing. The process is akin to funnelling from the wide to the narrow. Miles and Huberman (1984) suggest that careful data display (e.g. in graphics and diagrams) is an important element of data reduction and selection.

On the other hand Gibbs (2007, p. 4) argues that qualitative data analysis, far from reducing data, actually increases its 'bulk, density and complexity' as it creates more texts such as notes, reflections, memos, summaries, reflexive insights and further notes in its attempt to generate thick descriptions, i.e. data which describe events in context plus participants' intentions, strategies and agency.

Geertz (1973, pp. 10–21) argues that thick descriptions include reflections on meanings attributed to situations and phenomena, turning a witnessed, momentary event into a written discourse which can be perused repeatedly and 'read' in different ways.

Qualitative data present several challenges. First, data are so rich that analysis involves selecting and ordering on the part of the researcher. As a result, this might involve some personal bias to which the researcher needs to be alert. Second, since the data obtained are all couched in 'social events', reporting involves a double *hermeneutic process* (Giddens, 1976) by which the researcher *interprets* the data from participants who have already interpreted their world, and then relates them to the audience in his/her own words. Further, the researcher is part of the world that he or she is researching. Hence the reporting and analysis should strive to catch the different definitions of the situation from the different participants, and to combine *etic* and *emic* analysis. *Emic* analysis focuses on the participants' own subjective interpretations and perceptions of the situation, whilst *etic* analysis focuses on objective analysis or external frameworks. The researcher's own analysis might be subject to criticism of lack of objectivity, though this can be attenuated by the researchers' reflexivity. Qualitative data analysis is often written in the first person, and with colloquial language rather than the conventional third person, passive voice and past tense used in much research.

In selecting, organizing, analysing, interpreting and reporting data and findings the researcher is faced with several decisions and issues. For example, there is a risk that, since data and interpretation are unavoidably combined, the subjective views of the researcher might lead to him or her being over-selective, unrepresentative and unfair to the situation in hand in the choice of data and the interpretation placed on them. Fact and interpretation are inseparable. As the post-positivists remind us, there is no theory-free observation, and the selection of which events and data to include are under the control of the researcher. Indeed as participants (including the researcher) act on the basis of their interpretations, those interpretations may, themselves, become facts in the situation, i.e. an interpretation can constitute a fact or data, and, in that constructed sense, the written accounts of them are themselves created interpretations (cf. Geertz, 1973, p. 14).

Ensuring validity and reliability in qualitative data analysis is challenging, as there may be few external points of appeal other than respondent validation. The researcher's choice of which data and events to include is almost inevitably personal, but this choice has to be fair to the phenomena under investigation and to the participants, and reflexivity is important here. This is echoed in the later edition of Whyte's (1993) Appendix A to his celebrated study of *Street Corner Society*, where he writes that:

> it seemed as if the academic world had imposed a conspiracy of silence regarding the personal experiences of field workers. … It was impossible to find realistic accounts that revealed the errors and confusions and the personal involvements that a field worker must experience. I decided to do my bit to fill this gap. In undertaking this task it seemed to me important to be as honest about myself as I could possibly be.
>
> (Whyte, 1993, pp. 358–9)

Further, he reports commentaries that the researcher:

> abandons any hope of establishing scientific conclusions, and speaks rather of 'rendering your account credible through rendering your person so'. … Ethnography takes … a rather introspective turn. To be a customary 'I-witness' one must, so it seems, first become a convincing 'I'. Ethnological writing thus comes to depend on persuasion of the reader. … I have come to recognize that the objective-subjective distinction is not as clear as I once thought. … We seek to observe behavior that is significant to our research purposes. Selection therefore depends upon some implicit or explicit theory – a process which is in large part subjective. But the choice is not random. If we specify our theoretical assumptions and the research methods we use, others can utilize the same assumptions and methods to either verify or challenge our conclusions.
>
> (Whyte, 1993, pp. 366–7)

Further, Whyte (p. 362) questions the practicality of, or necessity for, respondent validation, particularly if the researcher discovers something that might contradict or upset the values and practices of the group. There is 'the right of the researcher to publish conclusions and interpretations as he or she sees them' (p. 362). Respondent validation may be problematic as, for example, respondents:

- may change their minds as to what they wished to say, or meant, or meant to say but did not say, or wished to have included or made public;
- may have faulty memories and may have recalled events over-selectively, or incorrectly, or not at all;
- may disagree with the interpretations made by the researcher;
- may wish to withdraw comments made in light of subsequent events in their lives;
- may have said what they said in the heat of the moment or because of peer pressure or authority pressure;
- may feel embarrassed by, or nervous about, what they said.

If respondents are asked to validate the data, the data analysis and interpretation, then, as Gibbs (2007, p. 95) remarks, their responses become data. Respondents may wish to withdraw their comments, and they may be entitled to do this if informed consent was given for *all stages* of the research, but they may not be entitled to do this if the informed consent was given to participate in the research but not to alter the reporting, i.e. the researcher owns the data, once given. Respondents may wish to change them, or prevent their public disclosure. In these situations the researcher might wish to explore the reasons for this, which, in turn, can become part of the research.

Whilst many qualitative data derive from field notes, given the exigencies of the moment (the 'personal convenience' of the researcher (Hammersley and Atkinson, 1983, p. 173)) and the press for time, some of these may also use the researcher's own memory, and this might be fallible, selective and over-interpreting a situation (p. 172), i.e. 'there is no single correct way of retrieving the data for analysis' (p. 173).

Hence it is important not only to examine a situation and events through the eyes of the researcher, but also to use a range of data and to ensure that these data include the views of other participants in a situation, in order to give some 'externality' to the situation and to focus on actual things that happened which can be corroborated by other participants. The process is inductive and reflexive, yet true to the indicators and constructs of the interpretation made.

Hammersley and Atkinson (1983) indicate the importance of reflexivity in addressing validity and reliability in the analysis and writing-up of qualitative data (p. 173). They suggest that the qualitative data analysis itself becomes a text, i.e. a constructed interpretation, and that its organization, ordering, chronology chosen, selection of themes, and narrative style are subject to reflexivity (pp. 212–17). Hence, the validity of the selection, analysis and interpretation of events and the data included in analysis, whilst being inductively and reflexively chosen, and whilst being unavoidably personal and partly impressionistic, are not only that; they are also subject to the validity checks of having other participants' views included and a faithful record made of actual events which involve more than the single researcher.

Qualitative data can be analysed for their nomothetic properties (patterns and themes – both emergent and pre-ordinate – trends, commonalities, generalizations, similarities, laws of behaviour) and their idiographic properties (individual, unique events, people, behaviours, contexts, actions, intentions). Nomothetic approaches to data analysis are well represented in the work of Miles and Huberman (1994), whilst idiographic approaches are well represented in life histories, case studies, individual biographies, phenomenological research and narratives.

32.3 Ethics in qualitative data analysis

Qualitative data analysis frequently concerns individual cases and unique instances, and may involve personal and sensitive matters. This raises questions of identifiability, anonymity, confidentiality and privacy of individuals. Whilst numerical data can be aggregated so that individuals are not traceable, this may not be the case in qualitative data analysis, even if individuals are not named or are given pseudonyms. The researcher has an ethical obligation to address non-maleficence, loyalties (and to whom) and beneficence (see Chapter 7), and to ensure that the principle of *primum non nocere* is addressed: do no harm to participants. This may call for respondent validation and respondent clearance for what is included, which, in turn, places the researcher in a dilemma of whether to include material that has not been cleared or which participants indicate they do not wish to have included or with which they disagree (e.g. in the case of an interpretation).

Given that some qualitative data may be sensitive or personal, the researcher will need to consider not only who will perform any transcription, but also the ethical conditions (e.g. confidentiality) to which the transcriber must be subject. This extends to ensuring that data are kept securely, with appropriately restricted access, and researchers may need to check with software providers on the security of data and who has access.

Ethics here also engages issues of research integrity, consideration of the consequences of the research and its publication, ownership of the data and how it may be used, informed consent and disclosure.

We refer the reader to Chapters 7 and 8 of the present volume, in which ethical matters are discussed in detail.

32.4 Computer assisted qualitative data analysis (CAQDAS)

Software does not analyse material; humans do. Software, as Kelle (2004, p. 277), Gibbs *et al*. (2005) and Gibbs (2012) remark, organizes and structures data for subsequent analysis. Software processes material. There are many ways in which software can be utilized in supporting qualitative research: Computer Assisted Qualitative Data Analysis Software (CAQDAS) (Tesch, 1990; LeCompte and Preissle, 1993; Gibbs *et al*., 2005; Gibbs, 2007, 2012; Creswell, 2012; Denscombe, 2014; Marshall and Rossman, 2016; Paulus *et al*., 2017). As can be seen from the list below, its uses are diverse.

Bazeley and Jackson (2013) note that CAQDAS software (they refer to NVivo) is useful for managing data and ideas, querying and searching data, visualizing data and reporting from the data (p. 3). Data must be organized, managed, processed and stored, with easy access, and software is very useful for this. There are several CAQDAS packages for data processing (Flick, 2009, pp. 360–1; Lewins and Silver, 2009; Creswell, 2012; Gibbs, 2012; Marshall and Rossman, 2016; Paulus *et al*., 2017). From the early days of simple search and retrieval software, CAQDAS has developed into more powerful, flexible tools to help the researcher-as-analyst with all kinds of data, not only text-based. Software has several uses, for example:

- to organize data and files systematically, and to manage data files for a project/research, managing, storing and indexing data in an ordered and organized way, for example, by ascribing data to specific addresses and indexes (see Figure 32.1);
- to store data, notes and searches;
- to search and interrogate data and text;
- to make notes and edit, extend or revise them;
- to transcribe and annotate field notes and audio and visual data;
- to search and retrieve data from individual files or across data files, codes, notes, memos;
- to display data in different ways and to create visual data modelling and graphics;
- to display relationships of categories (e.g. hierarchical, temporal, relational, subsumptive, superordinate);
- to establish linkages between coding categories and to cross-check to see if data can be coded into more than one category, enabling linkages between categories and data to be discovered;
- to code data (i.e. words or very short phrases which describe the data in question, for later ordering, combining or retrieval) and to arrange codes into hierarchies (trees) and nodes (key codes), to enable preliminary coding of data to be undertaken, and to attach identification labels to units of text (e.g. questionnaire responses), so that subsequent sorting can be undertaken;
- to facilitate content analysis (e.g. frequencies of words, meanings, issues, themes, concepts, sequences, locations, people, etc.);
- to check data (e.g. proofread);
- to collate and segment data and make numerous copies of data;
- to enable memoing to take place, together with details of the circumstances in which the memos were written;

- to conduct a search for words or phrases in the data and to retrieve text;
- to annotate and append text to written, audio, graphic, image-based and visual data;
- to partition data into units which have been determined either by the researcher or in response to the natural language itself;
- to sort, re-sort, collate, classify and reclassify pieces of data to facilitate constant comparison and to refine schemas of classification;
- to code memos and bring them into the same schema of classification as that used for other data in the study;
- to assemble, re-assemble and recall data into categories;
- to undertake frequency counts (e.g. of words, phrases, codes);
- to establish the incidence of data that are contained in more than one category;
- to retrieve coded and noded data segments from subsets in order to compare and contrast data;
- to search for pieces of data which appear in a certain (e.g. chronological) sequence;
- to filter, assemble and relate data according to preferred criteria (e.g. words, codes, themes, issues, nodes);
- to link to external sources of data (e.g. Internet sites) at a single keystroke;
- to draw conclusions and to verify conclusions and hypotheses;
- to quote data in the final report;
- to generate and test theory;
- to export data into other formats/software;
- to communicate with other researchers or participants.

CAQDAS can import data files which contain text, graphics, audio, pictures, video (no sound), sound-and-video, numbers (e.g. Excel files) and which link to external sources such as Internet sites and other software (e.g. online survey websites) (see Figure 32.1); they can present data graphically in clusters, 'trees'

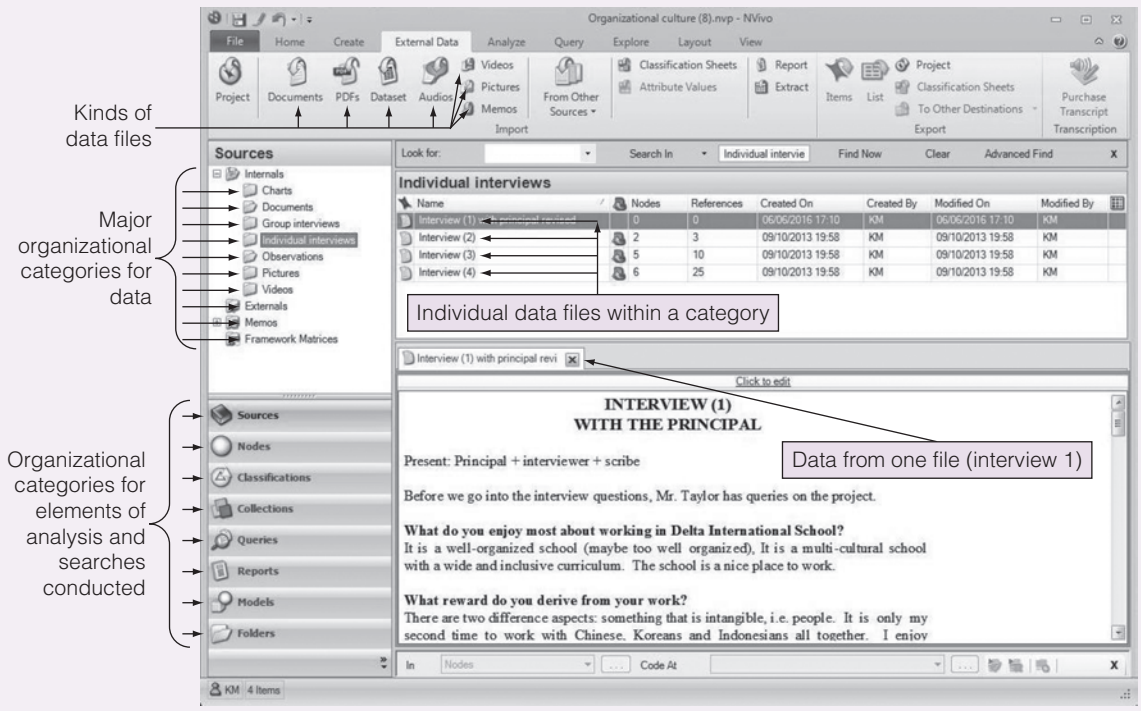


FIGURE 32.1 Organizing data in NVivo (Version 10)

Note

Screenshot reproduced with permission of NVivo qualitative data analysis Software; QSR International Pty Ltd. Version 10, 2012.

(hierarchies) and linkages, and they can assemble data from different sources.

Software for qualitative data analysis is useful for organizing data files. For example, in Figure 32.1, using NVivo, the data sources (see 'major organizational categories for data') have been set up by the researcher for a project on organizational culture, and are grouped as follows:

Internals: data files organized into categories and stored on the software for the project in question. The categories here are 'charts', 'documents', 'group interviews', 'individual interviews', 'observations', 'pictures', 'videos'. The top left of the screen in Figure 32.1 indicates the kinds of data files (documents, pdf files, data sets, audio, video, pictures and memos);
Externals: external website links that can be accessed for the project;
Memos: memos written by the researcher (Figure 32.2 lists these);
Framework Matrices: graphics of data presentation, with data alongside.

The lower left section of Figure 32.1 indicates where searches, queries, nodes, reports, models etc. prepared by the researcher are stored for ease of access and retrieval.

Figure 32.2 indicates the listing of memos in NVivo, and shows the text data for one of those memos (the memo entitled 'Observations on the pictures'), which is a memo on an observation of teachers' and senior managers' rooms in a secondary school. Figure 32.2 shows the actual contents of the memo (the text in the main central box of the screen shot), which can be stored along with the data file.

Many software programs enable the researcher to work with visual data, both still and moving images, and Figure 32.3 provides an example of annotating a photographic image of a teachers' working room in a secondary school. As with a text file, the image file here has the researcher's annotations stored along with the file. In Figure 32.3, working with NVivo, the researcher can see which part of the image is being referred to, as clicking onto the relevant row of annotated text also highlights that part of the image to

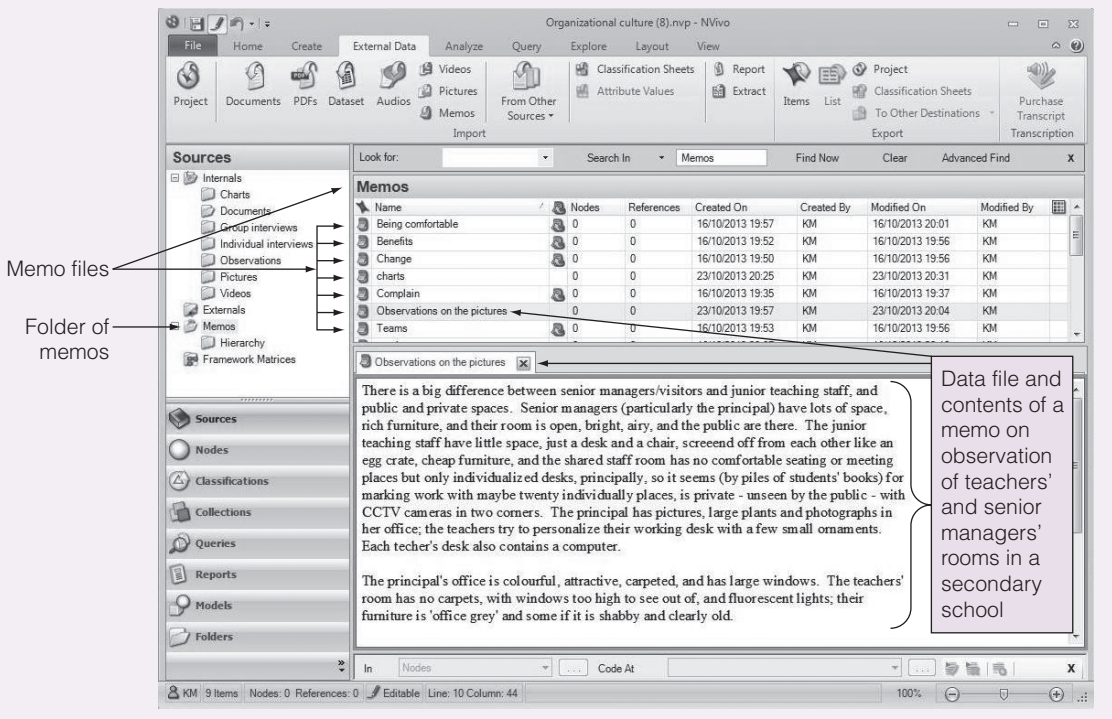


FIGURE 32.2 A sample memo on observation in NVivo (Version 10)

Note
Screenshot reproduced with permission of NVivo qualitative data analysis Software; QSR International Pty Ltd. Version 10, 2012.

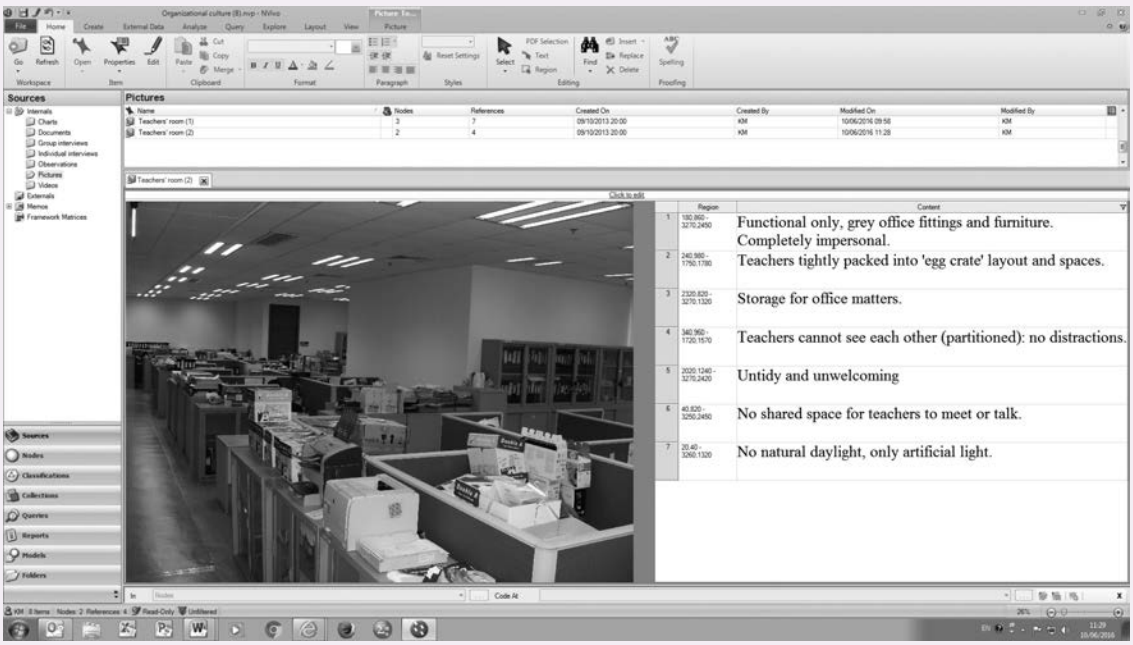


FIGURE 32.3 Annotated NVivo image file (Version 10)

Note

Screenshot reproduced with permission of NVivo qualitative data analysis Software; QSR International Pty Ltd. Version 10, 2012.

which reference is being made (this is not indicated in the figure).

CAQDAS software, then, can organize and store data in easily accessible and easily understood folders and files, and it can perform many other operations, for example:

- acting as a word processor (e.g. entering, editing and searching text);
- coding and retrieving many kinds of data (searching, summarizing, listing sequences of words, which enable data and texts to be split into smaller units and segments by relevant code and which list and organize and order codes and nodes). Coding data is part of a six-step sequence of data analysis (Kelle, 2000, p. 295): entering and formatting the text data; coding the data; memoing (with reference to specific segments of data); comparison of textual segments which have the same codes, to check for consistency; integrating the codes that have been generated, and memoing the codes; and developing the core category – a feature of grounded theory (see Chapter 37);
- managing data (e.g. searching, sorting and organizing);
- creating and presenting visual graphics such an networks of relationships;
- enabling theory building (e.g. through coding and the categorization and classification of codes and taxonomies to enable relations and superordinate and subordinate categories to be constructed);
- enabling conceptual networks to be plotted and visualized (graphic functions).

Which software one uses depends on the questions one wishes to ask of the data, the kinds of data one has, what one wishes to do with the data, the processes of analysis one wishes to conduct, the technical requirements of the software, the competence level of the researcher/user of the software, costs and the level of detailed required in the analysis.

Software is particularly effective at coping with the often-encountered problem of data overload and retrieval in qualitative research. Software enables the researcher to use codes, memos, hypertext systems, selective retrieval and co-occurring codes, and to perform quantitative counts of qualitative data types. In turn this enables linkages of elements to be created and networks to be built, and, ultimately, theory generation to be under-

taken. Software can assist in the generation of grounded theory through coding, constant comparison, linkages, memoing, annotations and appending, use of diagrams, verification and, ultimately, theory building. For a full discussion of coding and grounded theory we refer the reader to Chapters 34 and 37 respectively.

Kelle and Laurie (1995, p. 27) suggest that computer-aided methods can enhance: (a) validity (by the management of samples and data), and (b) reliability (by retrieving all the data on a given topic, thereby ensuring trustworthiness of the data) without losing contextual factors (cf. Gibbs, 2007, p. 106). An important feature here is the speed of organized and systematic data collation and retrieval; though data entry is time consuming, software can subsequently process and retrieve data rapidly.

There are several computer packages for processing qualitative data, for example: Anvil; AQUAD; ATLAS.ti; C-I-SAID; Dedoose; Diction; ELAN; ETHNOGRAPH; HyperRESEARCH; Kwaliton; Linguistic Inquiry; MAXQDA; NVivo; Qualrus; Quirkos; Transana. Widely used software is NVivo, ATLAS.ti and MAXQDA, though fitness for purpose is the key decision in deciding which software to use. We provide the websites of software packages and their evaluation in the companion website to this chapter, and Lewins and Silver (2009) also provide a guide on selecting software.

Gibbs (2007, 2012) focuses on three widely used packages: NVivo, MAXQDA and ATLAS.ti; Paulus *et al*. (2017) focus on NVivo and ATLAS.ti. These, like other software packages, share common features such as the ability to: (a) import, work with and display rich texts and multi-media material; (b) code text into key codes (nodes) and arrange codes and nodes into hierarchies and clusters; (c) sort, combine and retrieve text and data using different combinations and search strings/terms; (d) work with original documents/files using codes or combine selected extracts from documents/files using codes; (e) annotate, add memos, comments or additional documents to existing data files and documents; (f) sort material using codes; and (g) work with different kinds of data (textual, audio, images, videos, numbers, graphs, etc.). Additionally these programs can cope with large quantities of data rapidly and without any risk of human error in computation and retrieval, and they also release researchers from many mechanical tasks. With respect to words, phrases, codes, nodes and categories they can:

- search for and return data, text, terms, codes, nodes and categories, singly or in combination;
- filter text and data;
- return counts;
- present grouped data according to the selection criterion desired, both within and across data files;
- perform the qualitative equivalent of statistical analyses, such as:
  - Boolean searches (intersections of text which have been coded by more than one code or node, using ‘and’, ‘not’ and ‘or’, looking for overlaps and co-occurrences)
  - proximity searches (looking at clustering of data and related contextual data either side of, or near to, or preceding, or following, a node or code);
  - restrictions, trees, crosstabs (including and excluding documents for searching, looking for codes subsumed by a particular node, and looking for nodes which subsume others);
- construct dendrograms (tree structures) of related nodes and codes;
- present data in sequences and locate the text in surrounding material in order to provide the necessary context;
- locate and return similar passages of text or material (e.g. audio-visual);
- look for negative cases;
- look for terms in context (lexical searching);
- select text on combined criteria (e.g. joint occurrences, collocations);
- enable analyses of similarities, differences and relationships between texts and passages of text;
- annotate text and enable memos to be written about text.

Additionally, dictionaries and concordances of terms can be employed to facilitate coding, searching, retrieval and presentation.

Computer software can be particularly useful for searching, retrieving and grouping text, both in terms of specific words and in terms of words with similar meanings. Single words and word counts can overlook the importance of context, hence computer software packages have been developed that present Key-Words-In-Context. Most software packages have advanced functions for memoing, i.e. writing commentaries to accompany text that are not part of the original text but which may or may not be marked as material incorporated into the textual analysis. Software packages for qualitative data analysis typically include an annotation function, so that the researcher can annotate and append text, and the annotation is kept in the text but marked as an annotation.

CAQDAS does not do away with ‘the human touch’, as humans need to decide and generate the codes and categories, to verify and interpret the data; the software does

not generate the codes automatically. Similarly 'there are strict limits to algorithmic interpretations of texts' (Kelle, 2004, p. 277), as texts contain more than that which can be examined mechanistically and the software does not always suit the range and richness of analytic techniques associated with qualitative research (p. 283).

Woods *et al*. (2016) note that reflexivity remains a key feature of qualitative data analysis with software, and that this extends to consideration of the impact and influence of the software on the action, judgements and analysis conducted by researchers. Researchers cannot be unthinking software operatives, dredging for any information simply because software processes data, nor are they slaves to the software. On the one hand, CAQDAS can undermine reflexivity because: (a) the software does not compel researchers to adopt a particular philosophical position; (b) researchers can use the software relatively unthinkingly, without due reflection and without considering the acceptability of the values implicit in the software approach; (c) researchers can be forced in to a coding approach; and (d) CAQDAS can over-simplify complex issues (p. 393).

On the other hand, Woods *et al*. (2016) write that CAQDAS can encourage reflexivity if researchers think carefully about how the software might steer or influence their data preparation and analytical approaches, how it can improve and clarify their data analysis through otherwise difficult approaches (e.g. modelling and matrix construction). They also note that it can prompt analysis of data with which the researcher may be unfamiliar (e.g. visual data), i.e. it can develop new analytical skills in researchers, prompt them to see how to make appropriate use of these skills, and promote confidence in their analyses. Researchers using CAQDAS, just as those who do not, must be reflexive, and CAQDAS, like other approaches, can stimulate or inhibit reflexivity, depending on the user and his/her conscious decision making (p. 398).

Richards (2002) remarks on the tendency of some software packages to focus on 'code and retrieve' techniques (p. 266), with the risk that software encourages researchers to opt for coding and patterning to the neglect of more complex interrogation of texts (p. 269). Indeed, many researchers do not wish to use coding techniques with their qualitative data but are more concerned to review their texts iteratively (p. 270). Kelle (2004), Flick (2009) and Gibbs (2012) argue that software may be more closely aligned to the technique of grounded theory than to other techniques (e.g. hermeneutics, discourse analysis), that it may drive the analysis rather than *vice versa* (cf. Crowley *et al.*, 2002), and that it has a preoccupation with coding categories. Bazeley and Jackson (2013) note that CAQDAS: raises fears (some unfounded) that software use: 'can distance researchers from their data' (p. 7); privileges code and retrieve methods over other analytic strategies; renders analysis mechanistic and more positivist; and supports grounded theory approaches over other equally valid approaches (p. 7).

Gibbs (2012) and Gläser and Laudel (2013) raise the concern that if too great an emphasis is placed on coding and its applications then some important context may be stripped out of the data when they are assembled by codes alone. Gibbs (2007, 2012) reminds researchers that the use of software is only as good as the codes that have been used and the care taken with coding data, i.e. if poor codes or poor coding have been used and undertaken respectively (e.g. inconsistent coding, or coding that overlooks some text, or miscoding, or using a different code for the same kind or meaning of data) then poor results are likely to ensue (i.e. a problem of reliability). This applies similarly to searching codes, terms or combinations that have been undertaken.

Further, Flick (2009, p. 370) worries that the practicalities of data entry, coding and retrieval with software might distract researchers from the 'real' task of hermeneutically understanding, thinking about and explaining the meanings of the research and the texts. Indeed Taylor and Gibbs (2010) note that coding may be unsuitable for discourse analysis and narrative analysis (see also Chapter 35 here).

However, Bazeley and Jackson (2013) aver that these common criticisms misrepresent recent CAQDAS, as software is much more flexible than such criticisms suggest; it can remove the drudgery of some elements of qualitative data analysis without removing the creativity involved in it (p. 9) and, indeed, software does much more than simply code and retrieve. Indeed, Kelle (1997, para. 3.2) notes that ETHNOGRAPH is rooted in ethnographic and phenomenological research, that MAXQDA has its roots in Weberian 'ideal types' and AQUAD has its roots in Popperian methodology. Similarly Woods *et al*. (2016) note that ATLAS.ti was originally developed to support a grounded theory approach to data analysis.

Despite these answers to criticisms, software for qualitative data analysis does not give the same added value as that which one finds in quantitative data-analysis software (which automatically yields statistics), and textual or data input is a laborious process (Flick, 2009, p. 359). The 'added value' of CAQDAS software packages may not be as great as their statistical counterparts, for they require a significant amount of time and effort in preparing and entering transcribed and other word data and other kinds of data, which the software helps to search, organize, store, retrieve, link and collate. For statistical packages (e.g. SPSS) the

return on effort is much greater, as the software gives test results that do not have their simple equivalent in qualitative data-analysis software.

García-Horta and Guerra-Ramos (2009, p. 152) argue that qualitative software is no substitute for the requirement and capability of the researcher to 'assign meaning, identify similarities and differences, establish relations' between data. Indeed they suggest that, whilst software for qualitative data analysis might be useful for working with structure, software packages cannot currently handle the making of meaning, interpreting data, working out categories, making decisions on coding and interpreting the outcomes of analysis and processing (p. 153). However, advances have been on some fronts, for example in software for natural language processing (Crowston *et al*., 2012), and Paulus and Lester (2016) indicate how ATLAS.ti can be used for conversational and discourse analysis.

Paulus *et al*. (2017) suggest that, in reporting qualitative data analysis, researchers should move beyond conventional statements such as 'data were processed with NVivo' or 'coding and thematic analysis were undertaken with ATLAS.i' to include more information on how, exactly, the software was used and the steps that were taken in conducting the analysis with the software. This could include, for example, providing concrete details on:

- creating and storing text and other data;
- the sequence of the analysis using the software, and the key features of each step taken;
- annotating text;
- coding text and data;
- memo writing;
- linking data;
- creating themes;
- creating 'families' of data and data types, files, memos;
- merging data files;
- linking data to coding, memos, annotations etc.;
- creating visual graphics, hierarchies, networks, models, matrices, clustering, Boolean searches;
- exploring relationships;
- searching and retrieving material, for example, by quotations, words, texts, codes in single files and across files;
- conducting 'query' searches of data, for example, text searches, word frequency counts, coding, matrix coding;
- organizing and managing data;
- categorizing data;
- handling large-scale data (e.g. data reduction and display);
- indicating how the analysis was conducted with the software (e.g. narrative analysis, grounded theory approaches; searching for conceptual similarities and differences; locating quotes; word frequency counts);
- advantages and dangers of using the software (i.e. reflexivity);
- ensuring comprehensive and exhaustive data inclusion and usage;
- reflexivity and transparency;
- addressing reliability, validity, dependability, transferability and confirmability of the findings.

Paulus and Lester (2016) and Paulus *et al*. (2017) report how researchers using ATLAS.ti were able to document their decision making with regard to data analysis, demonstrating transparency, reflexivity, systematization and rigour. They also argue that it is preferable to use the active rather than the passive voice in writing up the analysis and findings, in order to avoid giving the impression that it was the software driving the analysis, rather than the researcher.

There are many websites that contain useful materials on qualitative data analysis and we identify these in the companion website for this chapter. Many of these provide links to a host of other websites providing guidance and resources for qualitative data analysis.

CAQDAS has moved great distances from simply searching and retrieving, and, as Paulus and Lester (2016) note, many of the criticisms of CAQDAS are based on outdated or incomplete understanding of what it can do and what is its potential for qualitative data analysis in a variety of traditions and approaches.

Companion Website

The companion website to the book provides data files and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections and a full set of word-based data files specifically prepared for NVivo, concerning a single project on assessment and testing (these have also been saved into Word documents). This resource can be found online at: **www.routledge.com/cw/cohen**.

Organizing and presenting qualitative data

CHAPTER 33

There are several ways in which qualitative data can be organized and presented. In this chapter we introduce some important, useful and widely used ways. These address several issues, including:

- tabulating data
- ten ways of organizing and presenting data analysis
- narrative and biographical approaches to data analysis
- systematic approaches to data analysis
- methodological tools for analysing qualitative data

We provide several worked examples here, for clarification. It is important for the researcher to index and provide a record of the provenance of the data, i.e. to record the dates, context, time, participants, researcher, location and so on, so that the setting for the data, and indeed their chronology, can be determined – the latter being useful in charting how situations emerge, evolve, change, lead to other situations, how networks emerge and how causality might be established. We outline several examples of data analysis and presentation in this chapter and the next.

33.1 Tabulating data

Tables are useful for data reduction and data display – key elements of qualitative data analysis as mentioned in Chapter 32 (Miles and Huberman, 1984, 1994). The following example illustrates simple summary and clear, tabulated data presentation and commentary. It derives from a doctoral thesis.

Example: Chinese children learning English – an example of analysing and presenting interview data

Here interview data are presented question by question. In what follows, where the data for respondents in each age phase are similar, they are grouped into a single set of responses by row; where there are dissimilar responses they are kept separate (Tables 33.1 to 33.4). The left-hand column in each table indicates the number of each participant in the research (1–12) and the level which the participant taught (e.g. P1, F3 etc.), so, for example, '1–3: P1' means the responses of participants 1–3, who taught Primary 1 classes; the right-hand column indicates the responses. In many cases, as can be seen, participants *all* gave similar responses in terms of the actual items mentioned and the coverage of items specified. A brief summary comment is provided after each table.

The data concern problems that school children experience in learning English in China. The data set reproduced is incomplete and has been selected for illustrative purposes only. Note that the data are not *verbatim*, but have already been summarized by the researcher, i.e. what is presented here is not the first stage of the data analysis, as the first stage was transcription.

The coding is as follows:

P1–P6 = Primary forms (1–6): P1 = year 1, P2 = year 2, etc.

F = Secondary forms (1–5): F1 = Form 1 (first year of secondary school), F2 = Form 2 (second year of secondary school, etc.)

The numbers preceding each letter in the left-hand column of these tables refers to the number ascribed to the teacher. There were twelve teachers in all, six from primary schools and six from secondary schools.

Table 33.1 indicates that English teaching and learning at school have not really achieved their intended purposes. Students: (a) are poor at understanding written or spoken English, speaking, reading, listening and writing; (b) have limited abilities in English, regardless of the number of years of learning it; (c) use low-level memorization which leads to superficial learning; (d) experience poor teaching and learning; (e) can enter university, even though their standard is poor, as there are many universities taking students; (f) do not require English to gain employment.

Comment: The Primary English teachers had a wider range of views than the Secondary teachers; there was greater unanimity between the Primary teachers in

TABLE 33.1 THE EFFECTIVENESS OF ENGLISH TEACHING

Q6: The effectiveness of English teaching

1–3: P1	■ Students neither understood written or spoken English nor were able to speak or write very well. ■ Though students started learning English at a very young age, their standard was still very low as they could not really understand or use English.
4–6: P6	■ Students could not speak, read or write English well. ■ Students had a low standard as they could not read, write or speak English. ■ They used memorization to learn and thus the English knowledge was very superficial and confined to limited vocabulary.
7–9: F3	■ On the whole, students' standard was low. English teaching and learning was not very successful. ■ Even with a poor knowledge of English students still managed to get jobs. ■ This was not an international city; English was not really that important even if students did not learn well.
10: F5	English teaching and learning were not very effective as students were not working hard and they resorted to memorization to learn English. However, students managed to get into universities.
11: F5	Students had learned at least some basic knowledge about English.
12: F5	It was effective to some extent as some students became English teachers themselves, having finished their university education.

comparison to the Secondary teachers; all the Form 3 Secondary teachers were unanimous in their comments; and all the Form 5 secondary teachers had different views.

Table 33.2 indicates that the strengths of English teaching were that: (a) students start to learn English very young; (b) schools had autonomy over the design of syllabuses. The weaknesses in English teaching were that: (a) insufficient emphasis was placed on understanding; (b) students were too young to learn English; (c) syllabuses were unrealistic in their demands, being too rich, leading teachers to a 'spoon-feeding' mentality in their teaching; (d) undue pressure was put on teachers and students because of the demands of the syllabus; (e) English had to compete with other languages for curriculum space. Hence students did not learn well, despite years of English lessons.

Comment: Apart from one Primary teacher, the other eleven teachers, drawn from both primary and secondary schools, were unanimous in the comments they gave.

Table 33.3 indicates that high class size (between thirty and fifty students, rising to sixty) and tight syllabuses exerted a significant impact on teaching methods and restrictions of class activities, because of class management issues. The nature of this influence was to adopt largely didactic and grammar-translation methods, with little extended recourse to using or 'thinking in' English. Teaching utilized some group activity, but this was very limited. Teachers used Chinese to explain English.

TABLE 33.2 THE STRENGTHS AND WEAKNESSES OF ENGLISH LANGUAGE TEACHING

Q7: Strengths and weaknesses of English language teaching

1: P1	Students started learning English at a very young age and they should be good at it. However, this could also be a disadvantage as students were too young to learn English and to understand what they were taught
2–6: P6 7–9: F3 10–12: F5	These respondents all commented that individual schools had great autonomy over syllabus design. Consequently, some syllabus contents were too rich to be covered within the limited time span. Therefore, it was hard to make adjustments, though students could not cope with the learning requirements. This put pressure on both teachers and students. Worse still, some schools made students learn other foreign languages apart from English, and that made the learning of English more difficult.

TABLE 33.3 TEACHING METHODS

Q9: Teaching methods

1–3: P1 4–6: P6 7–9: F3 10–12: F5	■ All respondents replied that teaching was mostly conducted on a didactic approach though they utilized visual aids and group activities to arouse students' interest, as they had a very tight syllabus to cover within the fixed number of periods. This method also gave them more control over the class, which was necessary as classes were usually big, between 30–50 and could rise to 60. ■ Whenever these teachers taught grammar, they relied heavily on the grammar-translation method. They used mostly Chinese (could be as much as 80%) to explain grammar, as that would make it easier for students to understand the explanation.

Comment: All the teachers here were unanimous in their comments, which fell mainly into two sets of points.

Table 33.4 indicates that students contributed significantly to their own success or failure in learning English. They: (a) were shy, afraid of making mistakes and of losing face; (b) had little interest in learning at all, let alone English; (c) were overloaded with other subjects, a situation exacerbated by their poor time management; (d) held negative attitudes towards the bookish nature of learning English and its unrelatedness to other curriculum subjects; (e) had too many other distractions; (f) had limited abilities in English; (g) had little incentive to learn fast, as they could repeat courses; (h) gave little priority to English; (i) had poor foundations for learning English; (j) had limited motivation or positive attitudes to learning English; (k) were given limited direction in their learning; (l) had limited incentive to learn English well, as universities required only a low standard of English.

Comment: There was a great variety of comments here. There were degrees of agreement: the teachers of the younger Primary children agreed with each other; the teachers of the older Primary children agreed, as did the teachers of the older Secondary children. The teachers of the younger Secondary children raised different points from each other. The four groups of teachers (younger Primary, older Primary, younger Secondary and older Secondary) raised different points from each other.

TABLE 33.4 STUDENT-RELATED FACTORS

Q11: Student-related factors

1–3: P1 4–6: P6	■ Students were shy and were afraid of 'losing face' when they made mistakes in front of the class. ■ Students basically had no interest in learning anything, especially a foreign language. ■ Students had too many subjects to learn, and learning English was too bookish. ■ There were too many other distractions such as surfing the Internet or going out with friends.
7: F3	■ Students could not relate learning English to other things they learned at school, so they had no interest. ■ Students' language learning ability was poor and they feared learning English. ■ Students were allowed to repeat programs, so they could become lazy and indifferent.
8: F3	■ Students spent too much time surfing the Net. ■ Students put more time into science rather than language subjects.
9: F3	■ Students' foundation was weak.
10–12: F5	■ Students lacked enthusiasm and 'proper' learning attitudes. ■ Students had poor time management. ■ Students were afraid of 'losing face' when they made mistakes in front of the class. They were shy as well. ■ Students had no direction in their learning and they had no plan for their future. Therefore, they did not learn well, especially a foreign language. ■ Students had many opportunities to enter universities, despite having a low standard of English.

For an example of the layout of tabulated word-based data and supporting analysis see the accompanying website.

Summary of the interview data

The issues that emerge from the interview data are striking in several ways. What characterizes the data is the widespread agreement of the respondents on the issues, for example:

1. There was absolute unanimity in the responses to questions 7 and 9.
2. There was very considerable, though not absolute, unanimity on question 11.
3. In addition to the unanimity observed in point (1), there was additional unanimity amongst the primary teachers in respect of question 11.
4. In addition to the considerable, though not absolute, unanimity observed in point (2), there was much unanimity amongst the primary teachers concerning question 6.

Such a degree of unanimity gives considerable power to the results, even though, because of the sampling used, they cannot be said to be representative of the wider population. However, the sample of experienced teachers was deliberately selected to provide an informed overview of key issues to be faced. It must be remembered that, though the unanimity is useful, the main purpose of the interview data was to identify key issues, regardless of unanimity, convergence or frequency of mention. The respondents articulated similar issues, however, and this signals that these may be important elements.

Further, the issues themselves are seen to lie in a huge diversity of fields, such that there is no single or simplistic set of problems or solutions. Hence, to complement the considerable unanimity of voice is a similar consensus in identifying the scope of the problem, yet the range of problems is vast. Both singly and together, the issues of English-language teaching, learning and achievement are complex. The messages are clear in respect of Form 5 students and their English teaching and learning:

i. English performance is weak in all its aspects – reading, writing, speaking, and listening – but it is particularly weak in speaking and writing.
ii. Local cultural factors exert an influence on learning English:
    - students do not wish to lose face in public (and the Chinese emphasis on gaining and maintaining face is powerful);
    - students are shy and afraid of making mistakes;
    - the pressure of examination success is universal and severe;
    - the local culture is not English; it is Chinese and there is little need for people to speak or use English.
iii. In some quarters, knowledge of English culture is seen to be an important element in learning English; this was refuted by the teachers in this sample.
iv. English is seen instrumentally, but this message has to be qualified, as many students gain employment and university entrance even though their English is weak. The fact that English is an international language has limited effect on student motivation or achievement.
v. Poor teaching and learning are significant contributors to poor performance, in several areas:
    - the emphasis on drill, rote learning and memorization;
    - the predominance of passive rather than active learning, with teaching as the delivery of facts rather than the promotion of learning and understanding;
    - the use of traditional didactic methods;
    - the reliance on a very limited range of teaching and learning styles;
    - the limited subject and pedagogical knowledge of English teachers, compounded by the lack of adequate initial and post-initial teacher education;
    - frequently the careful laying of foundations of English teaching and learning is absent;
    - students use so much Chinese during English lessons that they have little chance to think in English – they translate rather than think in English.

From the interview data it can be seen that the size of the problems and issues to be faced in English language teaching and learning is vast.

In this example, tables are carefully laid out to draw together similar sets of responses. The tables enable the researcher to see, at a glance, where similarities and differences lie between the two groups of respondents. Note also that after each table there is a summary of the main points to which the researcher draws attention, and that these comprise both substantive and overall comments (e.g. on the topic in hand and on the similarities and differences between the groups of respondents respectively). Finally, note that an overall summary of 'key messages' is provided at the end of all the tables and their commentaries. This is a very

abridged and selective example, and justice has not been done to the whole of the data that the original researcher used. Nevertheless the point clearly illustrates here that summarizing and presenting data in tabular form can address twin issues of qualitative research: data reduction through careful data display and commentary.

33.2 Ten ways of organizing and presenting data analysis

Organizing and presenting data (e.g. data display) are key issues in qualitative data analysis. Here we present ten ways of organizing and presenting analysis: the first two methods are by *people*, the next two are by *issue* or *theme*, the fifth is by *instrument*, the sixth is by *case studies*, the seventh is by *narrative account(s)*, the eighth is by *events*, whilst the ninth keeps these events and puts them in a chronology, by *time sequence* and *time frame*. The final method is by *theoretical perspectives*, enabling the researcher to gain some theoretical purchase on the phenomena under investigation.

In analysing qualitative data, a major tension may arise from using contrasting holistic and fragmentary/atomistic modes of analysis. The example above, of teaching English in China, is clearly atomistic, breaking down the analysis into smaller sections and units. It could be argued that this violates the wholeness of the respondents' evidence, and there is some truth to this, though one has to ask whether this is a problem or not. Sectionalizing and fragmenting the analysis can make for easy reading. On the other hand, holistic approaches to qualitative data presentation can catch the wholeness of individuals and groups, and this can lead to a more narrative, almost case study or story style of reporting, with issues emerging as they arise during the narrative. Neither approach is better than the other; researchers need to decide how to present data with respect to their aims and intended readership. The approaches outlined below address both holistic and atomistic approaches to qualitative data analysis.

In presenting the qualitative data analysis, researchers can utilize graphics, tables, matrices and clustering (e.g. Marshall and Rossman, 2016). However, these are presentational devices rather than analytical devices. The methods set out below deliberately indicate alternatives to coding; there is a risk that qualitative data analysts almost automatically turn to coding, but, as we indicate in Chapters 34 to 37, coding is only useful when it is fit for purpose, and there are many instances where it is an encumbrance and not fit for purpose in qualitative data analysis.

1 Organizing, analysing and presenting data by groups of people

In the example of teaching English above, the data were organized and presented by respondents, in response to particular issues. Where the respondents said the same, they were organized by groups of respondents in relation to a given issue. The groups of respondents were also organized by their membership of different strata in a stratified sample – teachers of: younger primary children, older primary children, younger secondary children and older secondary children. This is *one* way of organizing a qualitative data analysis: by *groups*. The advantage of this method is that it groups the data and enables themes, patterns and similar to be seen at a glance. Whilst this is a useful method for summarizing similar responses, the collective responses of an individual participant are dispersed across many categories and groups of people, and the integrity and coherence of the individual respondent risks being lost to a collective summary. Further, this method is often used in relation to a single-instrument approach, otherwise it becomes unwieldy (e.g. trying to put together the data derived from qualitative questionnaires, interviews and observations could be very cumbersome in this approach). So, researchers may find it helpful to use this approach instrument by instrument.

2 Organizing, analysing and presenting data by individual people

Here the total responses of a single participant are presented, and then the analysis moves on to the next individual. This preserves the coherence and integrity of the individual's response and enables a whole picture of that person to be presented, which may be important for the researcher. On the other hand, unless researchers are only interested in individual responses, it often requires them then to put together the issues arising *across* the individuals (a second level of analysis) in order to look for themes, shared responses, patterns of response, agreement and disagreement, to compare individuals and issues that each of them has raised, i.e. to summarize the data.

Different participants in a situation may have different perspectives on that situation. This method preserves the integrity of each person's perspective on that situation, and this can enable the researcher to find similarities and differences between them. For example, in a piece of curriculum innovation the school principal may support it, the subject head may support it, but the front-line teacher may disagree with it (too much work for little benefit), and the parent representative may be worried about it (e.g. will the students'

performance suffer?). This method enables these different views to be kept intact and attached to key players in the situation.

Whilst approaches that are concerned with people strive to be faithful to those involved, in terms of the completeness of the picture of them *qua* people, unless case study approaches are deemed to be driving the research, they are usually accompanied by a second round of analysis, which is of the issues that arise from the people, and it is to the matter of issues that we turn now.

3 Organizing, analysing and presenting data by issues or themes

Whilst this method is economical in making comparisons across respondents (the issue of data reduction through careful data display, mentioned earlier), again the wholeness, coherence and integrity of each individual respondent is lost.

The derivation of the issue/theme for which data are gathered needs to be clarified. For example, it could be that the issue has been decided *pre-ordinately*, in advance of the data collection. Then all the relevant data for that issue are simply collected together into that single basket: the issue in question. Whilst this is an economical approach to handling, summarizing and presenting data, it raises three main concerns:

- the integrity and wholeness of each individual can be lost, such that comparisons across the whole picture from each individual is almost impossible;
- the data can become decontextualized. This may occur in two ways: first in terms of their place in the emerging sequence and content of the research, the interview or the questionnaire (e.g. some data may require an understanding of what preceded a particular comment or set of comments); and second in terms of the overall picture of the relatedness of the issues, as this approach can fragment the data into relatively discrete chunks, thereby losing their interconnectedness and internal coherence;
- having had its framework and areas of interest decided pre-ordinately, the analysis may be unresponsive to additional relevant factors that could emerge *responsively* in the data. It is akin to lowering a magnet onto data: the magnet picks up relevant data for the issue in question but it also leaves behind data not deemed relevant, and these data risk being lost. The researcher, therefore, has to trawl the residual data to see if there are other important issues that have emerged which have not been caught in the pre-ordinate selection of categories and issues for attention.

The researcher, therefore, has to be mindful of the strengths and weaknesses not only of pre-ordinate categorization (and, by implication, include responsive categorization), but must also decide whether it is or is not important to consider the whole set of responses of an individual, i.e. to decide whether the data analysis is driven by people/respondents or by issues.

4 Organizing, analysing and presenting data by research question

This is a very useful way of organizing data, as it draws together all the relevant data for the exact issue of concern to the researcher, and preserves the coherence of the material. It returns the reader to the driving concerns of the research, thereby 'closing the loop' on the research questions that, in many kinds of research, drive the inquiry. In this approach all the relevant data from various data streams (interviews, observations, questionnaires etc.) are collated to provide a collective answer to a research question. There is usually a degree of systematization here, in that, for example, the numerical data for a particular research question will be presented, followed by the qualitative data, or *vice versa*. This enables patterns, relationships, comparisons and qualifications across data types to be explored conveniently and clearly.

This approach is self-evidently limited to those kinds of research which have clear research questions. Some kinds of qualitative research, for example, ethnography and phenomenography, may not have such precise research questions, in which case this approach may be unsuitable.

5 Organizing, analysing and presenting data by data-collection instrument

This approach is often used in conjunction with another approach, for example, by issue or by people. Here the data from each instrument are presented, for example, all the interview data are presented and organized, then all the questionnaire data are presented, followed by all the documentary data and field notes, and so on. Whilst this approach retains fidelity to the coherence and integrity of the instrument and enables the reader to see clearly which data derive from which instrument, the instrument is often only a means to an end and further analysis will be required to analyse the *content* of the responses, for example, by issue and by people. Hence if it is important to know from which instrument the data are derived then this is a useful method; however, if that is not important then this could be adding an unnecessary level of analysis to the data. Further, connections between data could be lost if the data are presented instrument by instrument rather than across instruments.

6 Organizing, analysing and presenting data by case study or studies

Here organizing and writing up qualitative data is by one or more (e.g. a series of) case studies, or by combining case studies into an overall study that sets out common and singular features and properties of the cases (see also Miles and Huberman (1994) on within site and cross-site analysis). A series of individual case studies can be followed by an analysis that draws together common findings from the different case studies and also indicates the exclusive features of each. The researcher can also identify common themes in and across the case studies. Alternatively, if a theme has been decided in advance (pre-ordinately) or indeed responsively when reading through all the case studies (see content analysis and coding, discussed in Chapters 34 to 37), then materials from case studies can be used selectively to illustrate specific themes, whilst adhering to the principle of fidelity to the case in question. Whilst this approach keeps the richness of the data, it may not solve the common problem of data and detail overload, as each case study inevitably requires detailed reporting in order to be faithful to the detail of the context, person, causality etc.

The six methods above suggest a degree of systematization and coherence in analysing and presenting data. However, such coherence may not always obtain; life is messy and full of internal contradictions. Further, each situation can sustain multiple interpretations and perspectives, and participants in social situations may have differing views that are not susceptible to easy organization or singular, reductionist analysis. Methods 7 and 8 below can be used when such coherence and reductionism neither obtains nor is important, or, if it is important, in the interests of being faithful to the phenomenon and participants, to retain rich detail.

7 Organizing, analysing and presenting data by narrative(s)

This way of organizing the analysis is by constructing a narrative that may be in the form of a chronology, a logical analysis, a thematic analysis and a story or series of 'stories' from the research findings.

The celebrated work of Goffman (1963, 1968, 1969) provides outstanding examples of this approach. For example, his work *Asylums* (1968) provides narrative accounts of perspectives and lived experiences of patients and staff at a psychiatric hospital. The narratives of these two groups are markedly different, very rich in detail and succeed in catching the different perspectives of the participants. There are no coding, categorization, theoretical saturation, core category or other commonly advocated tools for qualitative data analysis (see Chapters 34 to 37), but the narrative accounts that he provides give readers insights into the lives and minds of participants that may not be yielded by more contrived approaches to qualitative data analysis. They catch multiple meanings and multiple stories that other approaches may not. We return to the work of Goffman in Chapter 35.

8 Organizing, analysing and presenting data by event

Here the *events* may or may not be in a time series. It goes almost without saying that the researcher must decide and disclose the criteria used to make the selection of the events. For example this may be in terms of critical incidents (Tripp, 1993, 1994), for example, those incidents which constitute a turning point in the lives of teachers, students, teaching, schooling etc. The event may be something as routine as a staff meeting, a parents' evening or an educational visit. Alternatively it might be something less common such as the decision to bring in a major curriculum reform or to change the assessment system.

The event in question may be a planned or an unplanned event, a typical or atypical incident, something which interrupts or reproduces the existing situation, a problem or a solution, a positive or negative event, a particularly meaningful or significant event for particular people, an event that expands our horizons or diminishes them, opens our minds or closes them and so on. The event may only become significant in retrospect, and this requires the researcher to be reflective and/or to catch the reflections of participants. Different people have different views of an event, and it may be useful to report such differing views.

9 Organizing, analysing and presenting data by time sequence and time frame

Here the researcher decides and justifies the appropriate overall time frame and the duration of each segment of time. This may lead to reporting an emerging narrative. The researcher may decide to have a standard unit of analysis, for example, a week, a month, half a term, a semester, a school year, or a flexible, mutable unit of analysis, changing to suit the events, which may link to the previous method – organizing, analysing and presenting the data by event. By having a fixed unit of analysis, for example, a week, a month, half a term, a semester, a school year, the researcher can see, for example, time when events were happening swiftly or slowly, whether much was happening or little was happening, whether there were periods of development, consolidation or stagnation. Importantly, organizing,

analysing and presenting data by time sequence and time frame can enable the researcher to address causality: cause and consequence.

10 Organizing, analysing and presenting data by theoretical perspectives

In this approach the researcher uses different theoretical lenses to examine and report on phenomena and emerging situations. For example, changes in the management of schools may be examined through the lenses of neo-liberalist market reforms, rational choice theory, new managerialist perspectives, economic theories, theories of change, leadership theories, complexity theory, sociological perspectives and so on. The researcher has to decide and defend the perspectives chosen (e.g. Goldthorpe, 2007), for example on the grounds of fitness for purpose, fidelity to the phenomena and explanatory potential. Chapter 6 presents a worked example of this, from Goldthorpe (2007), in which the author, in examining the 'persistent differentials in educational attainment' (p. 21) had to decide between Marxist theory, liberal theory, cultural theory and rational choice theory.

In looking at phenomena through the chosen theoretical lenses, the researcher must be aware that this may be selective, including, excluding and even distorting data. Imagine looking at a coloured picture wearing spectacles that block out the colour red; what we see is affected by this screening out.

This approach does not mean that the researcher has to decide on a single theoretical perspective. Indeed different participants may have different theoretical perspectives, and the researcher may feel it important to report these.

These ten ways are not all mutually exclusive, and they may be used in combination, so as to better answer the research purposes and questions.

33.3 Narrative and biographical approaches to data analysis

Narrative and biographical approaches are powerful ways of analysing and presenting qualitative data. Bruner (1986) remarks that humans make meaning in terms of 'storied text' which catch the human condition, human intentionality and the vividness of human experience very fully (pp. 14, 19) and the multiple perspectives and lived realities ('subjective landscapes') of participants (p. 29). They model the world (p. 7), starting with metaphors and metamorphosing into empirical statements by verifiable data. They make the familiar strange, 'rescue it from obviousness' (p. 24) and require the reader to fill in the gaps, i.e. they are an interactive medium (p. 24).

Stories personalize generalizations (Gibbs, 2007, p. 57) and are evidence-based. Further, they catch the chronology of events as they unfold over time, and this can enable the researcher to infer causality, coupled with the dramatic and dramaturgical power of carefully chosen words. Narratives can not only convey information but bring information to life. As the poet Pasternak remarks, events 'catch fire' on their way, through the reporting of personal experiences, dramatic events and even the simple unfolding of a sequence of activities, behaviours or people over time. Gibbs (2007, p. 60) comments that narratives not only pass on information but meet people's psychological needs in coping with life; they help a group to crystallize or define an issue, view, stance or perspective; they can persuade or create a positive image; they can help researchers and readers to understand the experiences of participants and cultures; and they can contribute to the structuring of identity (as indeed is the case with life histories and biographies). Narratives are a wonderful foil to the supremacy of coding and coding-derived analysis.

Biographies, too, tend to follow a chronology, to report critical or key events and moments, to report key decisions and people, and to establish causality. Indeed, for their authors, they may even be restorative of broken identities or shattered futures (Gibbs, 2007, p. 67).

Narratives and biographies may have a chronology (but this is not a requirement, as some narratives are structured by logical relations or psychological coherence rather than chronology). They may have a beginning, a middle and an end, they may include critical moments and decisions, complicating factors, evaluation and outcomes (see Labov's (1972) characteristics of a narrative as having: (a) an abstract; (b) an orientation (context); (c) complicating actions (sequences of events that decide the course of the narrative); (d) evaluation (indicating the significance of the narrative and its main points); (e) resolution (outcomes); and (f) a coda (a rounding off of the narrative)).

Narratives and biographies cannot record all the events; rather a selective focus is adopted, based on the criteria that the researcher wishes to use. These may include: key decision points in the story or narrative; key, critical events, themes, behaviours, actions, decisions, people, points in the chronology; meaningful events to the participants; reconstruction of the case history (Flick, 2009, p. 347); key places; and key experiences. Once the researcher has identified the textual units in the biography or narrative, based on the criteria that are fit for the researcher's purpose, the researcher can then analyse and interpret the text for the meanings contained in it, develop working hypotheses to explain

what is taking place, check these hypotheses against the data and the remainder of the text, see the text as a whole rather than as discrete units, ensure that different interpretations of the text have been considered and that the interpretation(s) chosen are the most secure in terms of fidelity to the text.

Following these stages of text selection, analysis, interpretation and checking is the construction of the final narrative. This can be undertaken in several ways, for example:

- by temporal sequence (a chronology);
- by a sequence of causal relations;
- by key participants;
- by key actions;
- by emergent or key themes;
- by key issues and clusters of issues;
- by biographies of the participants;
- by critical or key events;
- by turning points in a life history or biography;
- by different perspectives;
- by key decision points;
- by key behaviours;
- by individual case studies or a collective analysis of the unfolding of events for many cases/participants over time.

In constructing a narrative analysis (as indeed in other forms of qualitative data analysis), the researcher can introduce verbatim quotations from participants where relevant and illuminative; these can add life to the narrative and often convey the point very expressively – without it being mediated or softened by the academic language of the researcher. It is important to keep quotations short enough to convey the main point without distortion or exclusion of relevant details and context, but not so long that the reader does not know what is the point of the quotation, i.e. having to perform an analysis of the data for herself/himself (Gibbs, 2007, p. 97). When using verbatim quotations from participants, it is often useful to accompany them with the researcher's interpretive commentary. Quotations are often chosen for their ability to crystallize or exemplify an issue really well, or typically, or extremely, and the researcher must decide whether to identify the person who said it (see the discussion of ethics in Chapters 7 and 32).

Narrative analysis, together with biographical data, can give the added dimension of realism, authenticity, humanity, personality, emotions, views and values in a situation, and the researcher must ensure that these are featured in the narratives constructed. By 'telling a story', a narrative account, case study or biography breaks with the strictures of coding and the risk of disembodied text that can too easily result from coding and retrieval exercises; it keeps text and context together, retains the integrity of people rather than fragmenting bits of them into common themes or codes, and enables evolving situations, causes and consequences to be charted. A narrative account enables events to 'catch fire' as they unfold. Narratives are powerful, human and integrated; truly qualitative.

33.4 Systematic approaches to data analysis

Qualitative data analysis can be very systematic. Becker and Geer (1960) indicate how this might proceed:

1. comparing different groups simultaneously and over time;
2. matching the responses given in interviews to observed behaviour;
3. analysing deviant and negative cases;
4. calculating frequencies of occurrences and responses;
5. assembling and providing sufficient data but keeping the separate raw data from analysis.

Qualitative data analysis here is inevitably interpretive, hence is less a completely accurate representation and more of a reflexive, reactive interaction between the researcher and the decontextualized data that are already interpretations of a social encounter. As mentioned earlier, the analysis is a construction of meaning rather than a complete reflection of reality and, in this, reflexivity is an important feature. The issue here is that the researcher brings to the data her own preconceptions, interests, biases, preferences, biography, background and agenda. As Walford (2001, p. 98) writes: 'all research is researching yourself'. In practical terms it means that the researcher may be selective in her focus, or that the research may be influenced by the subjective features of the researcher. Robson (1993, pp. 374–5) and Lincoln and Guba (1985, pp. 354–5) suggest that such subjective features can include and/or be subject to:

- data overload (humans may be unable to handle large amounts of data);
- first impressions (early data analysis may affect later data collection and analysis);
- availability of people (e.g. how representative these are and how to know if missing people and data might be important);
- information availability (easily accessible information may receive greater attention than hard-to-obtain data);

- positive instances (researchers may over-emphasize confirming data and under-emphasize disconfirming data);
- internal consistency (the unusual, unexpected or novel may be under-treated);
- uneven reliability (the researcher may overlook the fact that some sources are more reliable/unreliable than others);
- missing data (the issue for which there are incomplete data may be overlooked or neglected);
- revision of hypotheses (researchers may over-react or under-react to new data);
- confidence in judgement (researchers may have greater confidence than is tenable in their final judgements);
- co-occurrence may be mistaken for association;
- inconsistency (subsequent analyses of the same data may yield different results).

The issue here is that great caution and self-awareness must be exercised by the researcher in conducting qualitative data analysis, as the analysis and the findings may say more about the researcher than about the data. For example, it is the researcher who sets the codes and categories for analysis, be they pre-ordinate or responsive (decided in advance of or in response to the data analysis respectively). It is the researcher's agenda that drives the research and she who chooses the methodology.

As the researcher analyses data, she will have ideas, insights, comments and reflections to make on data. These can be noted down in memos, and indeed memos can become data themselves in the process of reflexivity (though they should be kept separate from the primary data themselves). Glaser (1978) and Robson (1993, p. 387) argue that memos are not data in themselves but help the process of data analysis; this is debatable: if reflexivity is part of the data-analysis process then memos may become legitimate (secondary) data in the process or journey of data analysis. Computer software for qualitative data analysis enables researchers to write a memo and attach it to a particular piece of datum. There is no single nature or format of a memo; it can include subjective thoughts about the data, with ideas, theories, reflections, comments, opinions, personal responses, suggestions for future and new lines of research, reminders, observations, evaluations, critiques, judgements, conclusions, explanations, considerations, implications, speculations, predictions, hunches, theories, connections, relationships between codes and categories, insights, and so on. Memos can be reflections on the past, present and the future, thereby beginning to examine the issue of causality. There is no required minimum or maximum length; memos should be dated not only for ease of reference but also for a marking of the development of the researcher as well as of the research. Chapter 37 discusses memos in greater detail.

Memos are an important part of the researcher's self-conscious reflection on the data; they have considerable potential to inform the data-collection, analysis and theorizing processes. They should be written whenever they strike the researcher as important – during and after analysis. They can be written any time; indeed some researchers deliberately carry recording methods with them wherever they go (pen and paper, electronic means, both written and audio) so that ideas that occur can be recorded before they are forgotten. They enable the researcher to comment and theorize on events, situations, behaviours and so on as they are being analysed, and can focus on observations, methodological and theoretical matters, or personal matters (cf. Gibbs, 2007, pp. 30–1).

We have discussed systematic qualitative data analysis in some detail in Chapter 25, including systematic tactics for generating meaning from transcribed data (Miles and Huberman, 1994), systematic content analysis (see also Chapter 34), and systematic procedures for phenomenologically analysing interview data. We advise readers to review the relevant material in that chapter together with the comments on content analysis in Chapter 34.

33.5 Methodological tools for analysing qualitative data

There are several procedural tools for analysing qualitative data. LeCompte and Preissle (1993, p. 253) present analytic induction, constant comparison, typological analysis and enumeration as valuable techniques for the qualitative researcher to use in analysing data and generating theory. Additionally coding is discussed in Chapter 34 and the tools of grounded theory are discussed in Chapter 37.

Analytic induction is a term and process that was introduced by Znaniecki (1934) in deliberate opposition to statistical methods of data analysis. LeCompte and Preissle (1993, p. 254) suggest that the process is akin to the several steps set out above, in that: (a) data are scanned to generate categories of phenomena; (b) relationships between these categories are sought; (c) working typologies and summaries are written on the basis of the data examined; (d) these are then refined by subsequent cases and analysis; (e) negative and discrepant cases are deliberately sought to modify, enlarge or restrict the original explanation/theory. Denzin (1970, p. 192) uses the term 'analytical induction' to describe the broad strategy and sequence of participant observation that is set out below:

1. A rough definition of the phenomenon to be explained is formulated.
2. A hypothetical explanation of that phenomenon is formulated.
3. One case is studied in the light of the hypothesis, with the object of determining whether or not the hypothesis fits the facts in that case.
4. If the hypothesis does not fit the facts, either the hypothesis is reformulated or the phenomenon to be explained is redefined, so that the case is excluded.
5. Practical certainty may be attained after a small number of cases has been examined, but the discovery of negative cases disproves the explanation and requires a reformulation.
6. This procedure of examining cases, redefining the phenomenon, and reformulating the hypothesis is continued until a universal relationship is established, each negative case calling for a redefinition or a reformulation.

A more deliberate seeking of disconfirming (negative) cases is advocated by Bogdan and Biklen (1992, p. 72). Here the researcher searches for cases which do not fit the other data, or cases, or that do not fit expected patterns of findings. They can be used to extend, expand or modify the existing or emerging hypothesis. Bogdan and Biklen also enumerate five main stages in analytic induction:

1. In the early stages of the research a rough definition and explanation of the particular phenomenon is developed.
2. This definition and explanation is examined in the light of the data that are being collected during the research.
3. If the definition and/or explanation that have been generated need modification in the light of new data (e.g. if the data do not fit the explanation or definition) then this is undertaken.
4. A deliberate attempt is made to find cases that may not fit into the explanation or definition.
5. The process of redefinition and reformulation is repeated until the explanation is reached that embraces all the data, and until a generalized relationship has been established, which will also embrace the negative cases.

In *constant comparison* the researcher compares newly acquired data with existing data and categories and theories that have been devised and which are emerging, striving to achieve a perfect fit between these and the data. Hence negative cases or data which challenge these existing categories or theories lead to their modification until they fully accommodate all the data. We discuss this technique more fully in Chapters 34 and 37.

Typological analysis is essentially a classificatory process (LeCompte and Preissle, 1993, p. 257) wherein data are put into groups, subsets or categories on the basis of some clear criterion (e.g. acts, behaviour, meanings, nature of participation, relationships, settings, activities). It is the process of *secondary coding* (Miles and Huberman, 1984, 1994) where descriptive codes are then drawn together and put into subsets. Typologies are a set of phenomena that represent subtypes of a more general set or category (Lofland, 1970). Lazarsfeld and Barton (1951) suggest that a typology can be developed in terms of an underlying dimension or key characteristic. In creating typologies Lofland (1970) insists that the researcher must: (a) deliberately assemble all the data on how a participant addresses a particular issue: what strategies are being employed; (b) disaggregate and separate out the variations between the ranges of instances of strategies; (c) classify these into sets and subsets; and (d) present them in an ordered, named and numbered way for the reader.

The process of *enumeration* is one in which categories and the frequencies of codes, units of analysis, terms, words or ideas are counted. This enables incidences to be recorded and, indeed statistical analysis of the frequencies to be undertaken (e.g. Monge and Contractor, 2003; Johnson and Black, 2012; Lee *et al*., 2015). This is a method used in some forms of content analysis, and we address this topic in the next chapter.

This chapter has suggested several approaches to analysing and presenting qualitative data. It should be read in conjunction with the comments on qualitative data analysis in Chapters 25, 34 and 37, as they complement each other.

Companion Website

The companion website to the book provides data files and additional material and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Coding and content analysis

CHAPTER 34

A key element of qualitative data analysis is data management. Qualitative data are often extensive, and careful data reduction is key. Too much reduction and the integrity and detail of the data are lost; too little and data overload and loss of clarity ensue. This chapter addresses data reduction with data quality through coding and content analysis. It indicates how coding works, what concerns it raises and how to address such concerns. It indicates different kinds of coding (e.g. open, analytic, axial, selective, theoretical) and how to code data and organize codes. The chapter uses the basis of coding in conjunction with content analysis, indicating what it is and how to conduct and report content analysis. It outlines an eleven-step process of conducting content analysis. Worked examples of coding and content analysis are provided and the chapter notes that both inductive and deductive approaches are important in qualitative data analysis.

This chapter addresses coding and content analysis in qualitative data analysis, both of these widely used, and it introduces key issues in coding and content analysis, including:

- coding
- concerns about coding
- what is content analysis?
- how does content analysis work?
- a worked example of content analysis
- reliability in content analysis

34.1 Introduction

One of the enduring problems of qualitative data analysis is the reduction of copious amounts of data to manageable and comprehensible proportions. Data reduction is a key element of qualitative analysis (Miles and Huberman, 1984, 1994), performed in a way that attempts to respect the *quality* of the qualitative data. One common procedure for achieving this is content analysis, a process by which the 'many words of texts are classified into much fewer categories' (Weber, 1990, p. 15), reducing the material in different ways (Flick, 1998, p. 192). Categories are derived from theoretical constructs or areas of interest devised in advance of the analysis (pre-ordinate categorization) or developed from the material itself (responsive categorization). Before we turn to content analysis, it is important to consider the matter of coding, and we address this below.

34.2 Coding

A major approach to qualitative data analysis is coding, for example, Strauss and Corbin (1990); Kelle (1995, pp. 62–104); Lonkila (1995); Gibbs (2007, pp. 38–55); Flick (2009, pp. 305–32); Creswell (2012); Marshall and Rossman (2016). There are several kinds of codes and we explore these below. Texts may be lightly coded or densely coded (e.g. where a single piece of text has several codes attached to it). A code is simply a name or label that the researcher gives to a piece of text which contains an idea or a piece of information (Miles and Huberman, 1994; Gläser and Laudel, 2013). Gibbs (2007, p. 38) catches the nature of a code neatly when he writes that the same code is given to an item of text that says the same thing or is about the same thing. Seidel and Kelle (1995) suggest that codes can denote a text, passage or fact, and can be used to construct data networks. Coding text means that data from non-textual sources (e.g. audio, visual images, videos, graphs, numerical files, charts and graphics) have had textual material (e.g. annotations, commentaries, notes, memos) added, and the coding works with and on that textual material.

Coding is the ascription of a category label to a piece of data, decided in advance or in response to the data that have been collected. The same piece of text may have more than one code ascribed to it, depending on the richness and contents of that piece of text. Coding is the process of breaking down segments of text data into smaller units (based on whatever criteria are relevant), and then examining, comparing, conceptualizing and categorizing the data. The researcher goes through the text, marking it with codes (labels) that describe that text. The code name might derive from the researcher's own creation, or it may derive from the

words used in the text or spoken by one of the participants in the transcribed data (e.g. if the participant remarks that she is bored with the science lesson, the code may be 'bored': a short term that catches the essence of the text in question).

Codes may be decided in advance, *ex ante*, preordinate from theory and/or from the research question, or, to be faithful to the data, they may be responsive to, and emerge from, the data. Whilst Glaser and Strauss (1967) advocate ignoring literature or theory and going straight into the data, Gläser and Laudel (2013) argue that, echoing the post-positivists, this is 'epistemically naïve' as observations are unavoidably theory-laden and theory orders and classifies how and what we see and look for (p. 72). In reality, the researcher often starts with some codes already decided or in mind and adds to, modifies and adjusts these in response to the data.

Coding enables the researcher to identify similar information. The researcher can search, retrieve and assemble the data in terms of those items that bear the same code. Codes can be regarded as an indexing or categorizing system, like the index in a book, which gives all the references to that index entry in the book, and the data can be stored under the same code, with an indexed entry for that code. A list of codes can be stored (e.g. in software such as NVivo), accompanied by information such as who coded the data, when the coding was undertaken and what the code means (Gibbs, 2007, p. 41). By coding the data the researcher is able to detect frequencies (which codes occur most commonly) and patterns (which codes occur together), and the researcher can retrieve all the data that have the same code, both within and across files.

Coding can be performed on many kinds of textual data (Gibbs, 2007, pp. 47–8), focusing on, for example: specific acts, conversations, reports, behaviours, events, interactions, activities, contexts, settings, conditions, actions, strategies, practices, tactics, meanings, intentions, states, symbols, participation, relationships, constraints, causes, consequences and issues concerning the researcher's reflexivity. In short, nothing is ruled out.

Codes can be at different levels of specificity and generality when defining content and concepts. Some codes may subsume others, thereby creating a hierarchy of subordination and superordination, in effect creating a tree diagram of codes (and software can present this in a graphic). Some codes are very general; others are more specific. Codes are astringent, pulling together a wealth of material into some order and structure. They can maintain context specificity. Codes may be *descriptive* and might include (Bogdan and Biklen, 1992, pp. 167–72): situation codes; perspectives held by subjects; ways of thinking about people and objects; process codes; activity codes; event codes; strategy codes; relationship and social structure codes; methods codes. The researcher goes through the data ascribing codes to each piece of datum. A code is a word or abbreviation sufficiently close to that which it is describing for the researcher to see at a glance what it means (in this respect it is unlike a number). For example, the code 'trust' might refer to a person's trustworthiness; the code 'power' might refer to the status or power of the person in the group. This enables meanings to be seen at a glance, memorized and recalled easily.

Miles and Huberman (1984, 1994) advise researchers to keep codes as discrete as possible and to start coding earlier rather than later as late coding enfeebles the analysis, though there is a risk that early coding might influence too strongly any later codes. It is possible, they suggest, for as many as ninety codes to be held in the working memory whilst going through data, though clearly there is a back-and-forth process whereby some codes that are used in the early stages of coding might be modified subsequently and *vice versa*, necessitating the researcher to go through a data set more than once to ensure consistency, refinement, modification and exhaustiveness of coding (some codes might become redundant whilst others might need to be broken down into finer codes). Data might be recoded on a second or third reading, as codes that were used early on might have to be refined in light of codes that are used later, either to make them more discriminating or to conflate codes that are unnecessarily specific. Codes, they argue, should enable the researcher to catch the complexity and comprehensiveness of the data. They are derived through the dual processes of induction and deduction (1994, p. 111) and should be verifiable by data (p. 108).

Before coding the text, it is important to read and re-read it to obtain a thorough understanding of meanings and key issues, a sense of the entire text and immediate main ideas in it (cf. Creswell, 2012, p. 244). Then, in coding a piece of text, the researcher goes through the data systematically, typically line by line, and writes a descriptive code by the side of each piece of datum, for example:

Text	*Code*
The students will undertake problem solving in science	PROB
I prefer to teach mixed ability classes	MIXABIL

Here the codes are abbreviations, and this is common, enabling the researcher to understand immediately the

issue that they denote because they resemble that issue (problem solving and mixed ability teaching), rather than, for example, ascribing a number as a code for each piece of datum, where the number provides no clue as to what the datum or category concerns. Where they are not abbreviations, Miles and Huberman (1994) suggest that the coding label should bear sufficient resemblance to the original data so that the researcher can know at a glance, by looking at the code, what the original piece of datum concerned. We give a worked example of a coding exercise later in this chapter.

There are several computer packages that can help the researcher here (e.g. MAXQDA, ATLAS.ti, NVivo, ETHNOGRAPH), and they require the original transcript to be entered onto the computer. Software can also enable coded text from across several files to be collated into a single file. Figure 34.1 shows one example of this, where coded text using the code for organizational culture ('orgcult') from three files has been collated into a single file in NVivo, with the names of the original files included and the text which has been selected and coded from each file.

It is important for codes to be applied consistently, so that relevant data are coded consistently, no data are excluded and the same code is used. This enables retrieval, categorization, collation and separation of data (particularly if software is being used).

Often, in the first coding attempt, many new codes are generated; the subtlety of difference of codes may be unclear as the researcher goes further through the text, or the earlier codes may turn out to be unhelpful (e.g. too general), or the later codes may be too strongly influenced (or driven) by the earlier codes, or later coding may make the researcher feel that she or he wishes to alter the earlier coding, or there may be duplication or overlap of codes (e.g. the same kind of meaning but given slightly different codes), or there may be redundant codes (e.g. codes that only appear once or twice and which are more fittingly replaced by other codes in light of the remainder of the text). The point here is that coding is not a 'one-off' exercise; it requires reading and re-reading, assigning and reassigning codes, placing and replacing codes, refining codes and coded data; the process requires the researcher to

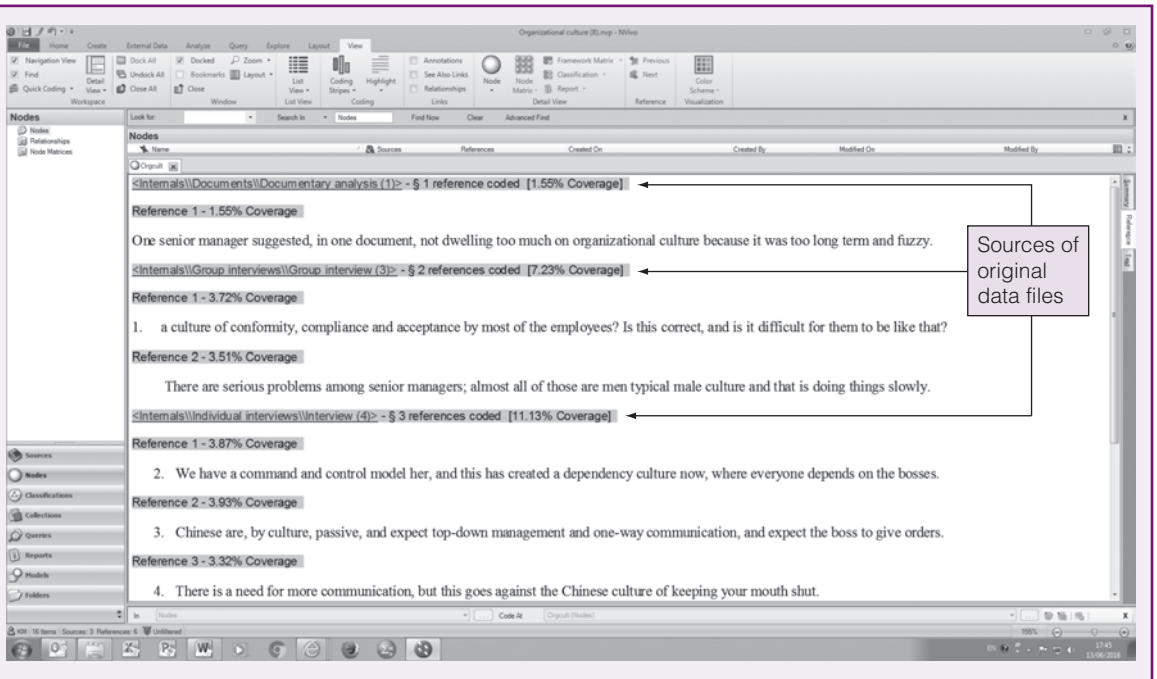


FIGURE 34.1 NVivo (Version 10) coded text for the code on organiztional culture, from several files collated into a single file

Note
Screenshot reproduced with permission of NVivo qualitative data analysis Software; QSR International Pty Ltd. Version 10, 2012.

go back-and-forth through the data on maybe several occasions, to ensure consistency and coverage of codes and data. Once the initial coding has been undertaken and checked then emergent themes, frequencies of codes, patterns of combinations of codes, key points, similarities and differences, variations and so on can be detected, and we discuss these in this chapter and the next.

Coding, argues Flick (2009, p. 310), can address fundamental questions such as 'who', 'why' 'what', 'where', 'how', 'when', 'how long', 'how much', 'how strong', 'what for' and 'by which'. These, he suggests, are useful questions in steering the coding exercise, particularly for open coding (discussed below).

There are different kinds of code: an open code, an analytic code, an axial code, a selective code and a theoretical code; we discuss these below. Though there is a suggestion in what follows that there is a sequence in coding, and indeed Strauss and Corbin (1990) suggest a sequence of three stages: open coding to axial coding to selective coding, this need not be the case, as different codes operate at different levels, and these are not necessarily driven by a pre-arranged sequence (Flick, 2009, p. 307).

Open coding

An open code is simply a new label that the researcher attaches to a piece of text to describe and categorize that piece of text (Strauss and Corbin, 1990, chapter 5). Open coding generates categories and defines their properties (the characteristics or attributes of a category or phenomenon) and dimensions (the location of a property along a given continuum) (Strauss and Corbin, 1990, p. 69). The authors give an example of the category/code 'colour', which has properties of hue, shade and intensity (p. 70). These properties, in turn, have dimensions: hue can be light to dark; shade can be light to dark; and intensity from high to low. Each category can have several properties, each of which has its own dimensional continuum (p. 70). The authors give an example of properties and dimensions for the category/code/label 'watching' (p. 72), such as: (a) property: 'frequency'; dimension: often to never; (b) property: 'extent'; dimension: more to less; (c) property: 'intensity': dimension: high to low; (d) property: 'duration'; dimension: long to short.

Open coding can be performed on a line-by-line, phrase-by-phrase, sentence-by-sentence, paragraph-by-paragraph, unit-of-text-by-unit-of-text (e.g. section) basis or a semantic unit. Then the codes can be grouped into categories, giving the categories a title or name, based on criteria that the researcher decides (e.g. concerning a specific theme, based on similar words, similar concepts, similar meanings etc.). The title of the category should be more abstract than the specific concepts or contents of the codes that it subsumes (Strauss and Corbin, 1990, p. 69). In undertaking such grouping, it is important that all the data fit into the group consistently, that there are no negative cases.

Open coding is usually the earliest, initial form of coding undertaken by the researcher.

Analytic coding

As its name suggests, an analytic code is more than a descriptive code. It is more interpretive. For example, whereas 'experimenting', 'controlling variables', 'testing' and 'measuring' are descriptive codes (e.g. in describing science activities), an analytic code here could be 'working like a scientist', 'doing science' or 'active science'; it draws together and gives more explanatory and analytic meaning to a group of descriptive codes. An analytic code might derive from the theme or topic of the literature or, responsively, from the data themselves.

Another example is where the descriptive codes given to teacher behaviour might be 'ignores disruption' (when a teacher ignores disruptive behaviour), 'interested students' (when a teacher only concentrates on those students who are interested in the lesson contents) and 'no response' (when a teacher does not respond to students shouting in class). The category might be 'teacher behaviour' and the analytic – more inferential – code might be 'teacher resignation' or 'teacher denial'.

Axial coding

An axial code is a category label ascribed to a group of open codes whose referents (the phenomena being described) are similar in meaning (e.g. concern the same concept). Axial coding is that set of procedures which the researcher follows whereby the data that were originally segmented into small units or fractions of a whole text through open coding are recombined in new ways (Strauss and Corbin, 1990, p. 96). An axial code refers to:

- *causal conditions*: events, activities, behaviours or incidents that lead to the occurrence of a phenomenon (p. 100);
- *a phenomenon*: an event, idea, activity, action, behaviour etc. (p. 100);
- *context*: a specific set of properties or conditions that obtain in a phenomenon, action or interaction (p. 101);
- *intervening conditions*: the broad, general conditions that have a bearing on the action or interaction in question (p. 103);

- *actions and interactions*: purposeful, goal-oriented processes, strategies or behaviours obtaining in an action (p. 104);
- *consequences*: outcomes for people, events, places etc., which may or may not have been predicted and which, in turn, may become the causes or conditions of further actions and interactions (p. 106).

For a worked example of these six areas, we refer readers to Buckley and Waring (2009), in which they diagrammatize the six areas and insert relevant data into them in their study of physical activity in children.

Axial coding connects related codes and sub-categories into a larger category of common meaning that is shared by the group of codes in question (thereby creating a hierarchy in which some codes are subsumed into the large axial category); an axial code, as its name suggests, is a category or axis around which several codes revolve.

Axial coding works within one category; it makes connections between sub-groups of that category and between one category and another. This might be in terms of the phenomena being studied, the causal conditions that lead to the phenomena, the context of the phenomena and their intervening conditions, and the actions and interactions of, and consequences for, the actors in situations.

Selective coding

Selective coding identifies the core categories of text data and integrates them to form a theory. It is the process of identifying the core category in a text, i.e. that central category or phenomenon around which all the other categories identified and created are integrated (Strauss and Corbin, 1990, p. 116), to which other categories are systematically related and by which it is validated. The authors argue that a selective code is very similar to an axial code, except that it is at a greater level of abstraction than an axial code. Creating the selective code requires: (a) a deep understanding of the main 'story line' (p. 117) (the descriptive overview of the main phenomenon being described and analysed, and its salient features); then (b) creating the core category; then (c) relating categories at the level of the dimensions identified; then (d) validating those relations in terms of the data that gave rise to them; and then (e) filling in any gaps in categories (pp. 116–17) to ensure the 'conceptual density' (p. 141) of the category, based on data collected. Though set out in a linear sequence, the authors indicate that, in fact, the process is iterative, and researchers move back and forth between steps (a) to (e).

Once codes have been assigned, ordered and grouped, they can be structured into hierarchies of subsumption, in which lower-order (e.g. descriptive) codes are subsumed under analytic and axial codes, which in turn are subsumed under a selective code. Hierarchies order codes and keep them tidy (Gibbs, 2007, p. 75), and indeed the creation of a hierarchy is itself part of data analysis, as the researcher ascribes meanings to the data. This is a pre-eminent function of CAQDAS software, in the creation of nodes, node trees and hierarchies.

The advice from Gibbs (2007, p. 77) is to keep hierarchies 'shallow' rather than 'deep', i.e. not too many levels. It is important, too, to ensure that the data contained in each code at each level are consistent with each other, hence the researcher has to constantly check and make comparisons across the data (the 'constant comparison' of grounded theory) (Glaser and Strauss, 1967) to ensure that they all fit together, with no exceptions or disconfirming data. Gibbs (2007, pp. 78–83) suggests that this can be done easily with tabulated data, and Chapter 33 provides examples of this, where data in columns can be compared or data in rows can be compared, to look for consistency, patterns, commonalities, relationships, similarities and differences (e.g. Tables 33.1 to 33.4). In such tabulated data (often where individuals are the rows and the issue is the column, it is possible to examine and compare individual cases (the rows) and different interpretations of the issues (the columns), for example, as shown in Table 34.1 (with fictitious data from a primary school, with the codes of 'AttSci' for attitudes to science lessons and 'AttMus' for attitudes to music lessons).

In this example, looking across the rows we can see that the children have positive attitudes but their interest is thwarted by distractions, lack of 'voice' and level of demand in the lessons; they all prefer practical work but this is not always done. One child seems to be more accommodating to the teacher's decisions than the other two, and one seems to be much less accommodating, i.e. there is variation on this dimension. Looking down the columns, we see very different attitudes within and between the two lessons. The table enables comparisons to be made, looking for similarities, differences, consistencies and inconsistencies, variations and homogeneity of responses, and deviant and extreme cases (cf. Gibbs, 2007, p. 96).

Theoretical coding

In theoretical coding, researchers see how codes and categories are integrated and fit together to create a theory or hypothesis. Here theoretical codes are the 'underlying logics' (Thornberg, 2012a, p. 89) that come

TABLE 34.1 TABULATED DATA FOR COMPARATIVE ANALYSIS

Name	*Attitudes to science lessons* *Code: AttSci.*	*Attitudes to music lessons* *Code: AttMus.*
Jane	Finds them difficult, but interesting. Too much homework which is not addressed in the class. Enjoys experiments but is not very good at them.	Enjoys listening to music, but there is too much singing to be done in class, and not enough playing or practical activity. The teacher only concentrates on those who are in the school choir.
John	Cannot concentrate because he finds the work boring and too 'bookish'. Prefers experiments but never has the chance to do them.	We are never allowed to choose the music to listen to, and the teacher's music is boring and old. Why do we have to use babyish instruments?
Stephen	Thoroughly enjoys the practical activities and the idea of exploring what went wrong in the experiments, and why.	I liked it when we were making up our own tunes in groups, but the class was very noisy. I don't like singing. I wish we were taught how to read and write proper music. Lots of children just 'mess around' in the music lesson, and that's horrible.

from pre-existing or emergent theories, together with the core category: that which has the greatest explanatory potential and to which the other categories and sub-categories relate most closely, repeatedly and consistently (discussed below and in Chapter 37). Glaser (1978) identifies 'families' of theoretical codes, including *the six Cs* (causes, contexts, contingencies, consequences, co-variances and conditions); *processes* (phases, progressions, passages, transitions, careers, trajectories, sequences, cycles); *type* (styles, classes, genre); *identity* (self-image, self-concept, self-worth, self-evaluation, identity, transformations of self); *degrees* (range, gradations, levels, limits); *culture* (social values, beliefs and norms). However, these 'families' may not exhaust theoretical possibilities, and the combining of categories may suggest other theoretical codes.

Coding is only an initial stage in qualitative data analysis; from initial coding the research can group codes into categories and then identify themes, trends and patterns (if indeed such trends and patterns exist), relations between themes, clusters of themes and issues, similarities and differences between themes and between data, and on to theory generation. This progression requires data from several texts to be addressed and combined which, in turn, can validate and increase the reliability of the findings and conclusions drawn. This can move towards 'saturation' (discussed below), wherein no additional data add to the key issues and findings.

34.3 Concerns about coding

The use of coding is governed by fitness for purpose; it is not suitable for all kinds of qualitative data analysis. Though coding is a central feature in many forms of qualitative data analysis, researchers need to ensure that it is the most appropriate way to analyse the data, as there is a risk of losing temporality, context and sequence in the coding and retrieval of text. For example, there is a temptation, perhaps to ascribe the same code to an observed behaviour regardless of the setting, the time (e.g. in a longitudinal study or a study that involves observation over several weeks), the prevalent conditions, states of mind, actors involved, intervening events, and so on, when, in fact the meaning and significance of the behaviour is not the same in different contexts or points in time.

Concerns have been raised that coding risks stripping out important contexts from the study and fragmenting holistic data into small segments, thereby losing the whole picture and having only a series of decontextualized codes (St Pierre and Roulston, 2006, p. 677; Blikstad-Balas, 2016, p. 9). In this case, the researcher may wish to write a narrative account rather than to abstract data from the several contexts in which they are set.

Coding can swamp the researcher with too many codes and may not reduce the data very much because the original textual material is still present which may contain irrelevancies (Gläser and Laudel, 2013). Further, St Pierre and Jackson (2014) raise the concern that coding treats words as 'brute data waiting to be

coded, labelled with other brute words (and even counted)' (p. 715), whereas qualitative data analysis is a much more humanistic and holistic activity. They argue against this 'vacuum cleaner approach' (p. 715), wherein all data are swept up and treated as equally important. Qualitative data analysis, they aver, should concern itself with the quality of the data and what is relevant and less relevant (e.g. informed by research questions and theory), with the researcher having to judge which words are and are not relevant for coding and to recognize that data are theory-laden, not theory-free.

St Pierre and Jackson (2014) also warn researchers against the propensity of humans to look for patterns where none exist, and that coding too easily feeds this fallacy. Further they argue that coding risks sacrificing an adequate theoretical or conceptual foundation to superficial presentation of data and codes which are supposed to 'speak for themselves' (p. 716). Rather, they comment that researchers must recognize 'the entanglement of research problems, concepts, emotions, transcripts, memories and images' (p. 717) that make up qualitative data analysis. There are no methodological steps, they argue, that one can simply drop unthinkingly onto data or that one can plan in advance; rather, they claim, qualitative data analysis needs to respond to the data, the research question and purposes, the conceptual and theoretical fields, the participants and so on, i.e. researchers must take a richer decision on data analysis rather than rushing headlong and thoughtlessly into coding.

Adair and Pastori (2011) suggest that qualitative data analysis, including coding, 'necessitates' conversation, debate and sophisticated, thoughtful decision making (p. 32). Texts, they note, are replete with 'subtlety, contradictions, metaphors, redundancy, and emotion', 'deeper meanings' and participants' voices (p. 33) which may not be susceptible to simplistic coding. Rather, coding should catch such subtlety and adopt both an *emic* approach (i.e. in terms that are meaningful to the participants) and an *etic* approach (objective, outsider descriptions), and *a priori* (pre-ordinate) and *a posteriori* (emergent) coding in order to catch such deeper meanings. In this respect, coding is just the start of the deeper process of qualitative data analysis, and not its terminus.

We return to coding in Chapter 37, as it is integral to grounded theory.

34.4 What is content analysis?

Having introduced coding, we are now in a position to consider 'content analysis'. The term is often used sloppily. In effect, it simply defines the process of summarizing and reporting written data – the main contents of data and their messages. 'Qualitative content analysis' (Gläser and Laudel, 2013) defines a strict and systematic set of procedures for the rigorous analysis, examination, replication, inference and verification of the contents of written data (Flick, 1998, p. 192; Krippendorp, 2004, p. 18; Mayring, 2004, p. 266). Texts are defined as any written communicative materials which are intended to be read, interpreted and understood by people other than the analysts (Krippendorp, 2004, p. 30). The intention of qualitative content analysis, argue Gläser and Laudel (2013), is to deliberately move from the original text to analysis of the information extracted from it (p. 13), focusing on the meanings of texts and their constituent parts.

Newby (2010, p. 485) reports three kinds of content analysis: 'conventional content analysis' (from coding); 'directed content analysis' wherein the coding structure derives from pre-existing theory or hypotheses; and 'summative content analysis' wherein keywords are selected based on previous research or the researcher's research interests. This is an entrée into the field, for, in reality, there are many kinds of content analysis, and we address these below.

Originally deriving from analysis of mass media and public speeches, the use of content analysis has spread to examination of any form of communicative material, both structured and unstructured. It may be used for those issues and problems which involve points of contact between culture and the social structure or to study social groups and interaction (Weber, 1990, p. 11). Content analysis can be undertaken with any written material, from documents to interview transcriptions, from media products to personal interviews. It is often used to analyse large quantities of text, facilitated by the systematic, rule-governed nature of content analysis, for example, as enabled by computer assisted analysis. It often uses categorization as an essential feature in reducing large quantities of data (Flick, 2009, p. 323).

Content analysis has several attractions. It is an unobtrusive technique (Krippendorp, 2004, p. 40). It focuses on language and linguistic features, meaning in context; it is systematic and verifiable (e.g. in its use of codes and categories) as the rules for analysis are explicit, transparent and public (Mayring, 2004, pp. 267–9). Further, as the data are in a permanent form (texts), verification through re-analysis and replication is possible.

Many researchers see content analysis as an alternative to numerical analysis of qualitative data. But this is not so, although it is widely used as a device for

extracting numerical data from word-based data. Indeed Anderson and Arsenault (1998, pp. 101–2) suggest that content analysis can describe the relative frequency and importance of certain topics as well as evaluate bias, prejudice or propaganda in print materials. Weber (1990, p. 9) sees the purposes of content analysis as including: (a) coding of open-ended questions in surveys; (b) revelation of the focus of individual, group, institutional and societal matters; (c) description of patterns and trends in communicative content. The latter suggestion indicates the role of statistical techniques in content analysis, indeed Weber (p. 10) suggests that the highest quality content-analytic studies use both quantitative and qualitative analysis of texts (texts defined as any form of written communication).

Content analysis takes texts and analyses, reduces and interrogates them into summary form through the use of both pre-existing categories and emergent themes in order to generate or test a theory. It uses systematic, replicable, observable and rule-governed forms of analysis in a theory-dependent system for the application of those categories. It can utilize coding, 'coding raw data into conceptually congruent categories' (Finfgeld-Connett, 2014, p. 342).

Krippendorp (2004, pp. 22–4) suggests that there are several features of texts that inform content analysis, including the fact that texts have no objective reader-independent qualities; rather they have multiple meanings and can sustain multiple readings and interpretations. There is no unitary meaning waiting to be discovered or described in them. Indeed, the meanings in texts may be personal and are located in specific contexts, discourses and purposes, and, hence, meanings have to be drawn in context. Content analysis, then: (a) describes the manifest characteristics of communication (p. 46) (asking who is saying what to whom, and how); (b) infers the antecedents of the communication (the reasons for, and purposes behind, the communication, and the context of communication) (Mayring, 2004, p. 267); and (c) infers the consequences of the communication (its effects). Krippendorp suggests (pp. 75–7) that content analysis is at its most successful when it can break down 'linguistically constituted facts' into four classes: attributions, social relationships, public behaviours and institutional realities.

34.5 How does content analysis work?

Ezzy (2002, p. 83) suggests that content analysis starts with a sample of texts, defines the units of analysis (e.g. words, sentences) and the categories to be used for analysis, reviews the texts in order to code them and place them into categories, and then counts and logs the occurrences of words, codes and categories. From here statistical analysis and quantitative methods are applied, leading to an interpretation of the results. Put simply, content analysis involves coding, categorizing (creating meaningful categories into which the units of analysis – words, phrases, sentences etc. – can be placed), comparing (categories and making links between them) and concluding – drawing theoretical conclusions from the text.

Anderson and Arsenault (1998, p. 102) indicate the quantitative nature of content analysis when they state that 'at its simplest level, content analysis involves counting concepts, words or occurrences in documents and reporting them in tabular form'. This succinct statement catches essential features of the process of content analysis:

- breaking down text into units of analysis;
- undertaking statistical analysis of the units;
- presenting the analysis in as economical a form as possible.

Denscombe (2014, pp. 283–4), echoing Anderson and Atsenault, sets out a six-stage process of content analysis:

1. Choosing an appropriate sample of data.
2. Breaking down text into smaller component units of analysis.
3. Developing appropriate categories for analysing the data.
4. Coding the units to fit the categories.
5. Conducting frequency counts of the occurrence of the units.
6. Analysing the text from the basis of the unit frequencies and how they relate to other units in the text.

Software can easily provide word frequency counts, and this can be useful. For example, in analysing inaugural speeches of high-profile people, the frequency of the word 'I' in the inaugural speech of 2,095 words by former American president Obama was only two (0.1 per cent), whilst for the word 'we' it was sixty-one (2.9 per cent). By contrast, in the inaugural speech of 1,520 words by former Australian prime minister Julia Gillard, the frequency of the word 'I' was 561 (37 per cent) whilst for the word 'we' it was thirteen (8.6 per cent). That sends signals, and many speech writers are sensitive to this; numbers are revealing. However, such an approach can mask some other important features of content analysis, including, for example, examination of the interconnectedness of units of analysis

(categories), the emergent nature of themes and the testing, development and generation of theory.

Flick (2009, p. 326) summarizes several stages of content analysis:

1. Defining the units of analysis.
2. Paraphrasing the relevant passages of text.
3. Defining the level of abstraction required of the paraphrasing.
4. Data reduction and deletion (e.g. removing paraphrases that duplicate meaning).
5. Data reduction by combining and integrating paraphrases at the level of abstraction required.
6. Putting together the new statements into a category system.
7. Reviewing the new category system against the original data.

More fully, the whole process of content analysis can follow several steps.

Step 1: define the research questions to be addressed by the content analysis

This includes what the researcher wants from the texts to be content-analysed. The research questions will be informed by, indeed may be derived from, the theory to be tested.

Step 2: define the population from which units of text are to be sampled

The population here refers not only to people but also, and mainly, to text – the domains of the analysis. For example, it could be newspapers and newspaper articles, programmes, interview transcripts, textbooks, conversations, public domain documents, journals, examination scripts, e-mails, online conversations etc.

Step 3: define the sample to be included

Here the rules for sampling people can apply equally well to documents. The researcher must decide whether to opt for a probability or non-probability sample of documents, a stratified sample (and, if so, the kind of strata to be used), random sampling, convenience sampling, domain sampling, cluster sampling, purposive sampling, systematic sampling, time sampling, snowball sampling and so on (see Chapter 12). Robson (1993, pp. 275–9) indicates the need for careful delineation of the sampling strategy here, for example, such-and-such a set of documents, or time frame (e.g. of newspapers), or television programmes, or interviews. Key issues in sampling people also apply to the sampling of texts: representativeness, access, size of the sample and generalizability of the results.

Krippendorp (2004, p. 145) indicates that there may be 'nested recording units', where one unit is nested within another, for example, with regard to newspapers that have been sampled: an item is nested in a paragraph, which is nested in an article, which is nested in an issue, which is nested in a particular newspaper (p. 145). This is the equivalent of stage sampling, discussed in Chapter 12.

Step 4: define the context of the generation of the document

This examines, for example: how the material was generated (Flick, 1998, p. 193); who was involved; who was present; where the documents came from; how the material was recorded and/or edited; whether the person was willing to, was able to and did tell the truth; whether the data were accurately reported (Robson 1993, p. 273) and corroborated; the authenticity and credibility of the documents; the context of the generation of the document; the selection and evaluation of the evidence contained in the document.

Step 5: define the units of analysis

This can be at very many levels, for example, a word, phrase, sentence, paragraph, whole text, people and themes. Robson (1993, p. 276) includes here, for newspaper analysis, the number of stories on a topic, column inches, size of headline, number of stories on a page, position of stories within a newspaper, the number and type of pictures. His suggestions indicate the careful thought that needs to go into the selection of the units of analysis. Different levels of analysis will raise different issues of reliability (discussed later). It is assumed that the units of analysis will be classifiable into the same category of text with the same or similar meaning in the context of the text itself (semantic validity) (Krippendorp, 2004, p. 296), though this can be problematic (discussed later). The description of units of analysis will also include the units of measurement and enumeration. The *coding unit* defines the smallest element of material that can be analysed, whilst the *contextual unit* defines the largest textual unit that may appear in a single category.

Krippendorp (2004, pp. 99–101) distinguishes three kinds of units. *Sampling units* are those units which are included in, or excluded from, an analysis; they are units of selection. *Recording/coding units* are units contained within sampling units, i.e. smaller than sampling units, thereby avoiding the complexity that characterizes sampling units; they are units of description. *Context units* are units of text which set boundaries on what is to be noted, i.e. the scope of the information which informs the coding of the material (p. 103).

Krippendorp continues by suggesting a further five kinds of sampling units: *physical* (e.g. time, place, size); *syntactical* (words, grammar, sentences, paragraphs, chapters, series etc.); *categorical* (members of a category have something in common); *propositional* (delineating particular constructions or propositions); and *thematic* (putting texts into themes and combinations of categories). The issue of categories signals the next step.

The criterion here is that each unit of analysis (category: conceptual, actual, classification element, cluster, issue) should be as discrete as possible whilst retaining fidelity to the integrity of the whole, i.e. each unit must be a fair rather than a distorted representation of the context and other data. The creation of units of analysis can be done by ascribing *codes* to the data, akin to the process of 'unitizing' (Lincoln and Guba, 1985, p. 203).

Step 6: decide the codes to be used in the analysis

Hammersley and Atkinson (1983, pp. 177–8) propose that the first activity here is to read and re-read the data to become thoroughly familiar with them, noting also any interesting patterns, any surprising, puzzling or unexpected features, any apparent inconsistencies or contradictions (e.g. between groups, within and between individuals and groups, between what people say and what they do). Then, having become familiar with the text, the process of coding can take place, following the principles and mechanics of coding as set out earlier in this chapter.

Step 7: construct the categories for analysis

Categories are the main groupings of constructs or key features of the text, showing links between units of analysis. For example, a text concerning teacher stress could have groupings such as 'causes of teacher stress', 'the nature of teacher stress', 'ways of coping with stress' and 'the effects of stress'. The researcher will have to decide whether to have mutually exclusive categories (preferable but difficult), how broad or narrow each category will be, the order or level of generality of a category (some categories may be very general and subsume other more specific categories, in which case analysis should only operate at the same level of each category rather than having the same analysis which combines and uses different levels of categories). Categories are inferred by the researcher, whereas specific words or units of analysis are less inferential; the more one moves towards inference, the more reliability may be compromised, and the more the researcher's agenda may impose itself on the data.

Categories must be exhaustive in order to address content validity; indeed Robson (1993, p. 277) argues that a content analysis – a system of categories – can include: subject matter; direction (how a matter is treated, e.g. positively or negatively); values; goals; method used to achieve goals; traits (characteristics used to describe people); actors (who is being discussed); authority (in whose name the statements are being made); location; conflict (sources and levels); and endings (how conflicts are resolved).

This stage of constructing categories is sometimes termed the creation of a 'domain analysis'. This involves grouping the units into domains, clusters, groups, patterns, themes and coherent sets to form domains. A domain is any symbolic category that includes other categories (Spradley, 1979, p. 100). The researcher can recode the data into domain codes, or review the codes used to see how they naturally fall into clusters, perhaps creating overarching codes for each cluster. Hammersley and Atkinson (1983) show how items can be assigned to more than one category, and, indeed, see this as desirable as it maintains the richness of the data. This is akin to the process of 'categorization' (Lincoln and Guba, 1985): putting 'unitized' data into categories of descriptive and inferential information. Unitization is the process of putting data into meaning units for analysis, examining data and identifying what those units are. A meaning unit is simply a piece of datum which the researcher considers to be important; it may be as small as a word or phrase, or as large as a paragraph, groups of paragraphs, or indeed a whole text, provided that it has meaning in itself.

Spradley (1979) suggests that establishing domains can be achieved by four analytic tasks: (a) selecting a sample of verbatim interview and field notes; (b) looking for the names of things; (c) identifying possible terms from the sample; and (d) searching through additional notes for other items to include. He identifies six steps to achieve these tasks: (i) select a single semantic relationship; (ii) prepare a domain analysis sheet; (iii) select a sample of statements from respondents; (iv) search for possible cover terms and include those which fit the semantic relationship identified; (v) formulate structural questions for each domain identified; (vi) list all the hypothesized domains. Domain analysis, then, strives to discover relationships between symbols (p. 157).

Like codes, categories can be at different levels of specificity and generality. Some categories are general and overarching; others are less so. Typically codes are much more specific than categories. This indicates the difference between *nodes* and *codes*. A code is a label for a piece of text; a node is a category into which

different codes fall or are collected. A node can be a concept, idea, process, group of people, place, or indeed any other grouping that the researcher wishes it to be; it is an organizing category. Whilst codes describe specific textual moments, nodes draw together codes into a categorical framework, making connections between coded segments and concepts. It is rather like saying that a text can be regarded as a book, with the chapters being the nodes and the paragraphs being the codes, or the contents page indicating the nodes and the index indicating the codes. Nodes can be related in several ways, for example: one concept can define another; they can be logically related; and they can be empirically related (found to accompany each other) (Krippendorp, 2004, p. 296).

The construction of codes and categories might steer too much the research and its findings, i.e. the researcher may enter too far into the research process. For example, a researcher examining the extra-curricular activities of a school might conclude that the benefits of these are to be found in non-cognitive and non-academic spheres rather than in academic spheres, but this may be fallacious. It could be that it was the codes and categories themselves rather than the data in the minds of the respondents that caused this separation of cognitive/academic spheres from the non-cognitive/non-academic, and that if the researcher had specifically asked about or established codes and categories which established the connection between the academic and non-academic, then he would have found more than he did. This is the danger of using codes and categories to predefine the data analysis.

Step 8: conduct the coding and categorizing of the data

Once the codes and categories have been decided, the analysis can be undertaken. This concerns the actual ascription of codes and categories to the text, as described earlier in this chapter. Mayring (2004, pp. 268–9) suggests that *summarizing content analysis* reduces the material to manageable proportions whilst maintaining fidelity to essential contents, and that *inductive category formation* proceeds through summarizing content analysis by inductively generating categories from the text material. This is in contrast to *explicit content analysis*, the opposite of summarizing content analysis, which seeks to add in further information in the search for intelligible text analysis and category location. The former reduces contextual detail, the latter retains it. *Structuring content analysis* filters out parts of the text in order to construct a cross-section of the material using specified pre-ordinate criteria.

It is important to decide whether to code simply for the existence or the incidence of the concept. This is important, as it would mean that, in the case of the former – existence – the frequency of a concept would be lost, and frequency may give an indication of the significance of a concept in the text. Further, the coding will need to decide whether it should code only the exact words or those with a similar meaning. The former will probably result in significant data loss, as words are not often repeated in comparison to the concepts that they signify; the latter may risk losing the nuanced sensitivity of particular words and phrases. Indeed some speechmakers may deliberately use ambiguous words or those with more than one meaning.

Having performed the first round of coding, the researcher is able to detect patterns and themes and begin to make generalizations (e.g. by counting the frequencies of codes). The researcher can also group codes into more general clusters, each with a code, i.e. begin the move towards factoring the data.

Perhaps the biggest problem concerns coding and scoring of open-ended questions. Two solutions are possible here. Even though a response is open-ended, an interviewer, for example, may pre-code her interview schedule so that while an interviewee is responding freely, the interviewer is assigning the content of her responses, or parts of it, to predetermined coding categories. Classifications of this kind may be developed during pilot studies. Gläser and Laudel (2013) are strong advocates of this pre-coded (*ex ante*) approach, with the codes deriving 'from a theoretically derived set of categories' (p. 14) and the categories deriving from the 'same theoretical framework that already has guided data collection' (p. 15) and which, nevertheless are open to modification. The categories, they aver, should include the material dimensions of the matters in hand, the time dimension and the causal dimension.

Alternatively, data may be post-coded. For example, having recorded the interviewee's response, either by summarizing it during or after the interview itself, or verbatim by recording, the researcher can conduct content analysis of it and apply one of the available scoring procedures – scaling, scoring, rank scoring, response counting, etc.

Extraction of meaning in qualitative data analysis requires identifying 'the category to which the information belongs' (p. 17), in which the unit of analysis is often a paragraph, and the same paragraph may fit more than one category. Gläser and Laudel (2013) are clear to distinguish their process from coding, in that the former extracts meaning rather than text, and is more strongly theory-driven (*ex ante*) and research-question-driven than coding, without 'forcing' data into categories.

Step 9: conduct the data analysis

Once the data have been coded and categorized, the researcher can count the frequency of each code or word in the text, and the number of words in each category. This is the process of retrieval, which may be in multiple modes, for example words, codes, nodes and categories. Some words may be in more than one category, for example where one category is an overarching category and another is a sub-category. To ensure reliability, Weber (1990, pp. 21–4) suggests that it is advisable at first to work on small samples of text rather than the whole text, to test out the coding and categorization, and make amendments where necessary. The complete texts should be analysed, as this preserves their semantic coherence.

Words and single codes on their own have limited power, and so it is important to move to associations between words and codes, i.e. to look at categories and relationships between categories. Establishing relationships and linkages between domains ensures that the richness and 'context-groundedness' of data are retained. Linkages can be found by identifying confirming cases, by seeking 'underlying associations' (LeCompte and Preissle, 1993, p. 246) and connections between data subsets.

Weber (1990, p. 54) suggests that it is preferable to retrieve text by categories rather than single words, as categories tend to retrieve more than single words, drawing on synonyms and conceptually close meanings. One can make category counts as well as word counts and specify at what level the counting can be conducted, for example, words, phrases, codes, categories and themes.

The implication here is that the frequency of words, codes, nodes and categories provides an indication of their significance. This may or may not be true, since subsequent mentions of a word or category may be difficult in certain texts (e.g. speeches). Frequency does not equal importance, and not saying something (withholding comment) may be as important as saying something. Content analysis only analyses what is present rather than what is missing or unsaid (Anderson and Arsenault, 1998, p. 104). Further, as the researcher moves through a piece of text, he or she: may replace nouns with pronouns; must be careful to avoid continuously raising the same issue as such redundancy can lead to unhelpful repetition; must be aware that some issues or topics may be more difficult than others to identify; and must be aware that if the text is short then this might 'inhibit reference to the theme' (Weber, 1990, p. 73).

The researcher can summarize the inferences from the text, look for patterns, regularities and relationships between segments of the text, and test hypotheses. The summarizing of categories and data is an explicit aim of statistical techniques, for these enable trends, frequencies, priorities and relationships to be calculated. There are several approaches and methods for data analysis, for example (Krippendorp, 2004, pp. 48–53):

- extrapolations (trends, patterns and differences);
- standards (evaluations and judgements);
- indices (e.g. of relationships, frequencies of occurrence and co-occurrence, number of favourable and unfavourable items);
- linguistic re-presentations.

Once frequencies have been calculated, statistical analysis can proceed, using, for example:

- factor analysis (to group the kinds of response);
- tabulation (of frequencies and percentages);
- crosstabulation (presenting a matrix where the words or codes are the column headings and the nominal variables, e.g. the newspaper, the year, the gender, are the row headings);
- correlation (to identify the strength and direction of association between words, between codes and between categories);
- graphical representation (e.g, to report the incidence of particular words, concepts, categories over time or over texts);
- regression (to determine the value of one variable/word/code/category in relationship to another): a form of association that gives exact values and the gradient or slope of the goodness-of-fit line of relationship – the regression line;
- multiple regression (to calculate the weighting of independent variables on a dependent variable);
- structural equation modelling (to determine the multiple directions of inferred causality and the weightings of different associations in a pathway analysis of causal relations);
- dendrograms (tree diagrams to show the relationship and connection between categories and codes, codes and nodes).

The calculation and presentation of statistics are discussed in Chapters 38 to 44. At this stage we note that what starts as qualitative data – words – can be converted into numerical data for analysis (though St Pierre and Jackson (2014) argue against frequency counts, suggesting that this betrays the qualitative nature of qualitative data).

If a less quantitative form of analysis is required then, for example, one can establish linkages and

relationships between concepts and categories, examining their strength and direction (how strongly they are associated and whether the association is positive or negative respectively). Robson (1993, p. 401) suggests that drawing conclusions from qualitative data can be undertaken by counting, patterning (noting recurrent themes or patterns), clustering (of people, issues, events etc. which have similar features), relating variables, building causal networks and relating findings to theoretical frameworks.

It is also useful to try to identify core categories (see the later discussion of grounded theory). A core category is that which has the greatest explanatory potential and to which the other categories and sub-categories seem to be repeatedly and closely related (Strauss, 1987, p. 11).

Whilst conducting qualitative data analysis using numerical approaches or paradigms might be criticized for being positivistic, one should note that one of the founders of grounded theory – Glaser – is on record (1996) as saying that not only did grounded theory develop out of a desire to apply a quantitative paradigm to qualitative data, but that paradigmal purity was unacceptable in the real world of qualitative data analysis, in which *fitness for purpose* should be the guide.

The process of analysis continues until 'saturation' is reached, i.e. when additional data do not add anything more to understanding and making meaning of the data (Finfgeld-Connett, 2014, p. 348).

Step 10: summarizing

By this stage the investigator will be in a position to write a summary of the main features of the situation researched so far. The summary will identify key factors, key issues, key concepts and key areas for subsequent investigation. It is a watershed stage during the data collection, as it pinpoints major themes, issues and problems that have arisen, so far, from the data (responsively) and suggests avenues for further investigation. The concepts used will be a combination of those derived from the data themselves and those inferred by the researcher (Hammersley and Atkinson, 1983, p. 178).

At this point, the researcher will have gone through a preliminary sequence of theory generation for qualitative data (Patton, 1980):

1. finding a focus for the research and analysis;
2. organizing, processing, ordering and checking data;
3. writing a qualitative description or analysis;
4. inductively developing categories, typologies, and labels;
5. analysing the categories to identify where further clarification and cross-clarification are needed;
6. expressing and typifying these categories through metaphors;
7. making inferences and speculations about relationships, causes and effects.

Bogdan and Biklen (1992, pp. 154–63) identify several important factors that researchers need to address at this stage, including: forcing oneself to take decisions that will focus and narrow the study and decide what kind of study it will be; developing analytical questions; using previous observational data to inform subsequent data collection; writing reflexive notes and memos about observations, ideas, what is being learned; trying out ideas with subjects; analysing relevant literature whilst conducting the field research; generating concepts, metaphors and analogies and visual devices to clarify the research (Miles and Huberman (1984, 1994) strongly advocate the graphic display of data as an economical means of reducing qualitative data. Such graphics might serve both to indicate causal relationships as well as simply to summarize data).

Step 11: making speculative inferences

This is an important stage, for it moves the research from description to inference. Here the researcher, on the basis of the evidence, posits some explanations for the situation, some key elements and possibly even their causes. It is the process of hypothesis generation or the setting of working hypotheses that feeds into theory generation.

The stage of theory generation is linked to grounded theory, and we turn to this in Chapter 37. Here we provide an example of content analysis that does not use statistical analysis but which nevertheless demonstrates the systematic approach to analysing data that is at the heart of content analysis.

34.6 A worked example of content analysis

In this example the researcher has already transcribed data concerning stress in the workplace from, let us say, a limited number of accounts and interviews with some teachers, and these have already been summarized into key points. Imagine that each account/interview has been written up onto a separate file (e.g. computer file), and now they are all being put together into a single data set for analysis. What we have are already-interpreted, rather than verbatim, data.

Stage 1: extract the interpretive comments that have been written on the data

By the side of each, a code/category/descriptor word has been inserted (in capital letters), i.e. the summary data have already been collected together into thirty-three summary sentences.

1 Stress is caused by deflated expectation, i.e. stress is caused by annoyance with other people not pulling their weight or not behaving as desired, or teachers letting themselves down. **CAUSE**
2 Stress is caused by having to make greater demands on personal time to meet professional concerns. No personal time/space as a cause of stress. Stress is caused by having to compromise one's plans/desires. **CAUSE**
3 Stress comes from having to manage several demands simultaneously, **CAUSE** but the very fact that they are simultaneous means that they can't be managed at once, so stress is built into the problem of coping – it's an insoluble situation. **NATURE**
4 Stress from one source brings additional stress which leads to loss of sleep – a sign that things are reaching a breaking point. **OUTCOME**
5 Stress is a function of the importance attached to activities/issues by the person involved. **NATURE** Stress is caused when one's own integrity/values are not only challenged but called into question. **CAUSE**
6 Stress comes from 'frustration' – frustration leads to stress leads to frustration leads to stress etc. – a vicious circle. **NATURE**
7 When the best-laid plans go wrong this can be stressful. **CAUSE**
8 The vicious circle of stress, inducing sleep irregularity, which in turn induces stress. **NATURE**
9 Reducing stress often works on symptoms rather than causes – it may be the only thing possible **CAUSE** given that the stressors will not go away, but it allows the stress to fester. **CAUSE**
10 The effects of stress are physical which, in turn, causes more stress – another vicious circle. **OUTCOMES**
11 Stress from lowering enthusiasm/commitment/aspiration/expectation. **CAUSE**
12 Pressure of work lowers aspiration which lowers stress. **CAUSE**
13 Stress reduction through companionship. **HANDLING**
14 Stress because of things out of one's control. **CAUSE**
15 Stress through handling troublesome students. **CAUSE**
16 Stress because of a failure of management/leadership. **CAUSE**
17 Stress through absence of fulfilment. **CAUSE**
18 Stress rarely happens on its own; it is usually in combination – like a rolling snowball, it is cumulative. **NATURE**
19 Stress through worsening professional conditions that are out of the control of the participant. **CAUSE** Stress through loss of control and autonomy. **CAUSE**
20 Stress through worsening professional conditions is exponential in its effects. **NATURE**
21 Stress is caused when professional standards are felt to be compromised. **CAUSE**
22 Stress because matters are not resolved. **CAUSE**
23 Stress through professional compromise which is out of an individual's control. **CAUSE**
24 The rate of stress is a function of its size – a big bomb causes instant damage. **NATURE**
25 Stress is caused by having no escape valve; it's bottled up and causes more stress, like a kettle with no escape valve, it will stress the metal and then blow up. **CAUSE**
26 Stress through overload and frustration – a loss of control. Stress occurs when people cannot control the circumstances with which they have to work. **CAUSE**
27 Stress through overload. **CAUSE**
28 Stress through seeing one's former work being undone by others' incompetence. **CAUSE**
29 Stress because nothing has been possible to reduce the level of stress. So, if the boil of stress is not lanced, it grows and grows. **CAUSE NATURE**
30 Handling stress through relaxation and exercise. **HANDLING**
31 Trying to relieve stress through self-damaging behaviour – taking alcohol and smoking. **HANDLING NATURE**
32 Stress is a function of the importance attached to activities by the participants involved. **NATURE**
33 The closer the relationship to people who cause stress, the greater the stress. **NATURE**

The data have been coded very coarsely into four main categories. It might have been possible to have coded the data far more specifically, for example, each specific cause has its code, and indeed one school of thought would argue that it is important to generate the specific codes first. One can code for words (and, thereafter, the frequency of words) or meanings – it is sometimes dangerous to go for words rather than meanings, as people say the same things in different ways

Stage 2: sort data into key headings/areas

The codes that have been used fall into four main areas:

a causes of stress
b nature of stress
c outcomes of stress
d handling stress

Stage 3: list the topics within each key area/heading and put frequencies in which items are mentioned

For each main area the relevant data are presented together, and a tally mark (/) is placed against the number of times that the issue has been mentioned by the teachers.

a Causes of stress
- Deflated expectation/aspiration /
- Annoyance /
- Others not pulling weight /
- Others letting themselves down /
- Professional demands, for example, troublesome students /
- Demands on personal time from professional tasks /
- Difficulties of the job /
- Loss of personal time and space /
- Compromising oneself /one's professional standards and integrity ///
- Plans go wrong /
- Stress itself causes more stress /
- Inability to reduce causes of stress /
- Lowering enthusiasm/commitment/aspiration /
- Pressure of work /
- Things out of one's control //
- Failure of management/leadership /
- Absence of fulfilment /
- Worsening professional conditions /
- Loss of control and autonomy //
- Inability to resolve situation /
- Having no escape valve /
- Overload at work /
- Seeing one's work undone by others /

b Nature of stress
- Stress is a function of the importance attached to activities/issues by the participants /
- Stress is inbuilt when too many simultaneous demands are made, i.e. it is insoluble /
- It is cumulative (like a snowball) until it reaches a breaking point /
- Stress is a vicious circle //
- The effects of stress are exponential /
- The rate of stress is a function of its size /
- If stress has no escape valve then that causes more stress //
- Handling stress can lead to self-damaging behaviour (smoking/alcohol) /
- The closer the relationship to people who cause stress, the greater the stress /

c Outcomes of stress
- Loss of sleep/physical reaction //
- Effects of stress themselves cause more stress /
- Self-damaging behaviour /

d Handling stress
- Physical action/exercise /
- Companionship /
- Alcohol and smoking /

Stage 4: go through the list generated in Stage 3 and put the issues into groups (avoiding category overlap)

Here the grouped data are re-analysed and re-presented according to possible groupings of issues under the four main headings (causes, nature, outcomes and handling of stress: (a)–(d) below).

a *Causes of stress*:

i Personal factors
- Deflated expectation/aspiration /
- Annoyance /
- Demands on personal time from professional tasks /
- Loss of personal time and space /
- Stress itself causes more stress /
- Inability to reduce causes of stress /
- Lowering enthusiasm/commitment/aspiration /
- Things out of one's control //
- Absence of fulfilment /
- Loss of control and autonomy //
- Inability to resolve situation /
- Having no escape valve /

ii Interpersonal factors
- Annoyance /
- Others not pulling weight /
- Others letting themselves down /
- Compromising oneself/one's professional standards and integrity ///
- Seeing one's work undone by others /

iii Management
- Pressure of work /
- Things out of one's control //
- Failure of management/leadership /
- Worsening professional conditions /
- Seeing one's work undone by others /

- **iv** Professional matters
  - Others not pulling weight /
  - Professional demands, for example, troublesome students /
  - Demands on personal time from professional tasks /
  - Difficulties of the job /
  - Compromising oneself/one's professional standards and integrity ///
  - Plans go wrong /
  - Pressure of work /
  - Worsening professional conditions /
  - Loss of control and autonomy //
  - Overload at work /
- **b** *Nature of stress*
  - **i** Objective
    - It is a function of the importance attached to activities issues by the participants /
    - Stress is inbuilt when too many simultaneous demands are made, i.e. it is insoluble /
    - It is cumulative (like a snowball) until it reaches a breaking point /
    - Stress is a vicious circle //
    - The effects of stress are exponential /
    - The rate of stress is a function of its size /
    - If stress has no escape valve then that causes more stress //
    - Handling stress can lead to self-damaging behaviour (smoking/alcohol) /
  - **ii** Subjective
    - Stress is a function of the importance attached to activities issues by the participants /
    - The closer the relationship to people who cause stress, the greater the stress /
- **c** *Outcomes of stress:*
  - **i** Physiological
    - Loss of sleep /
  - **ii** Physical
    - Physical reactions //
    - Increased smoking /
    - Increased alcohol /
  - **iii** Psychological
    - Annoyance /
- **d** *Handling stress*
  - **i** Physical
    - Physical action/exercise /
  - **ii** Social
    - Social solidarity, particularly with close people ///
    - Companionship /

Stage 5: comment on the groups or results in Stage 4 and review their messages

Once this stage has been completed, the researcher is then in a position to draw attention to general and specific points, for example:

1. There are very many causes of stress (give numbers).
2. There are very few outlets for stress, so it is inevitable, perhaps that stress will accumulate.
3. Causes of stress are more rooted in personal factors than any others – management, professional etc. (give frequencies here).
4. The demands of the job tend to cause less stress than other factors (e.g. management), i.e. people go into the job knowing what to expect, but the problem lies elsewhere, with management (give frequencies).
5. Loss of control is a significant factor (give frequencies).
6. Challenges to people and personal integrity/self-esteem are very stressful (give frequencies).
7. The nature of stress is complex, with several interacting components (give frequencies).
8. Stress is omnipresent.
9. Not dealing with stress compounds the problem; dealing with stress compounds the problem.
10. The subjective aspects of the nature of stress are as important as its objective nature (give frequencies).
11. The outcomes of stress tend to be personal rather than outside the person (e.g. systemic, or system-disturbing) (give frequencies).
12. The outcomes of stress are almost exclusively negative rather than positive (give frequencies).
13. The outcomes of stress tend to be felt non-cognitively, for example, emotionally and psychologically, rather than cognitively (give frequencies).
14. There are few ways of handling stress (frequencies), i.e. opportunities for stress reduction are limited.

The stages in this example illustrate several of the issues raised in the preceding discussion of content analysis, though the example here does not undertake word counts or statistical analysis, and, being fair to content analysis, this could – some would argue even 'should' – be a further kind of analysis. This analysis raises several issues:

- the researcher has looked *within* and *across* categories and groupings for patterns, themes and generalizations, as well as exceptions, unusual observations etc.;

- the researcher has had to decide whether frequencies are important, or whether an issue is important even if it is only mentioned once or a few times;
- the researcher has looked for, and reported, disconfirming as well as confirming evidence for statements;
- the final stage of the analysis is theory generation, to account for what is being explained about stress. It might also be important, in further analysis, to try to find causal relationships here: what causes what and the directions of causality; it may also be useful to construct diagrams (with arrows) to show the directions, strength and positive/negative nature of stress.

34.7 Reliability in content analysis

There are several issues to be addressed in considering the reliability of texts and their content analysis, indeed in analysing qualitative data using a variety of means, for example:

- Witting and unwitting evidence (Robson, 1993, p. 273): witting evidence is that which was intended to be imparted; unwitting evidence is that which can be inferred from the text, and which may not be intended by the imparter.
- The text may not have been written with the researcher in mind and may have been written for a very different purpose from that of the research (a common matter in documentary research); hence the researcher will need to know or be able to infer the intentions of the text.
- The documents may be limited, selective, partial, biased, non-neutral and incomplete because they were intended for a different purpose other than that of research (an issue of validity as well as of reliability).
- It may be difficult to infer the direction of causality in the documents – they may have been the cause or the consequence of a particular situation.
- Classification of text may be inconsistent (a problem sometimes mitigated by computer analysis), because of human error, coder variability (within and between coders) and ambiguity in the coding rules (Weber, 1990, p. 17).
- Texts may not be corroborated or be able to be corroborated.
- Words are inherently ambiguous and polyvalent (the problem of homographs), for example, what does the word 'school' mean? A building; a group of people; a particular movement of artists (e.g. the impressionist school); a department (a medical school); a noun; a verb (to drill, to induct, to educate, to train, to control, to attend an institution); a period of instructional time ('he stayed after school to play sports'); a modifier (e.g. a school day); a sphere of activity (e.g. 'the school of hard knocks'); a collection of people adhering to a particular set of principles (e.g. the utilitarian school); a style of life (e.g. 'a gentleman from the old school'); a group assembled for a particular purpose (e.g. a gambling school), and so on. This is a particular problem for computer programs which may analyse words devoid of their meaning.
- Coding and categorizing may lose the nuanced richness of specific words and their connotations.
- Category definitions and themes may be ambiguous, as they are inferential.
- Some words may be included in the same overall category but they may have more or less significance in that category (and a system of weighting the words may be unreliable).
- Words that are grouped together into a similar category may have different connotations and their usage may be more nuanced than the categories recognize.
- Categories may reflect the researcher's agenda and imposition of meaning more than the text may sustain or the producers of the text (e.g. interviewees) may have intended.
- Aggregation may compromise reliability. Whereas sentences, phrases and words and whole documents may have the highest reliability in analysis, paragraphs and larger but incomplete portions of text have lower reliability (Weber, 1990, p. 39).
- A document may deliberately exclude something for mention, overstate an issue or understate an issue.

At a wider level, the limits of content analysis are suggested by Ezzy (2002, p. 84), where he argues that, due to the pre-ordinate nature of some forms of coding and categorizing, content analysis is useful for testing or confirming a pre-existing theory rather than for building a new one, though this perhaps understates the ways in which content analysis can be used to generate new theory, not least through a grounded theory approach (see Chapter 37). In many cases content analysts know in advance what they are looking for in text, and perhaps what the categories for analysis will be. Ezzy (p. 85) suggests that this restricts the extent to which the analytical categories can be responsive to the data, thereby confining the data analysis to the agenda of the researcher rather than the 'other'. Indeed Finfgeld-Connett (2014) draws attention to the risk of 'verifying the obvious' (p. 342) by a deductive rather than an inductive approach to content analysis, and

Mayring (2004, p. 269) argues that if the research question is very open or if the study is exploratory, then more open procedures than content analysis may be preferable, for example, grounded theory.

Though inductive approaches may be ruled out of the early stages of content analysis, they may feature in the later stages, as themes and interpretations may emerge inductively from the data and the researcher, rather than only or necessarily from the categories or pre-existing theories themselves. Hence to suggest that content analysis denies induction or is confined to the testing of pre-existing theory (Ezzy, 2002, p. 85) is to misrepresent the flexibility of content analysis. Indeed Flick (1998) suggests that pre-existing categories may need to be modified if they do not fit the data. Both inductive and deductive approaches are important in qualitative data analysis.

Companion Website

The companion website to the book provides data files and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Discourses

CHAPTER 35

Conversations, narratives and autobiographies as texts

Whilst coding represents one major approach to analysing qualitative data (discussed in the previous chapters), nevertheless it is only one way. In this chapter we provide very different methods of analysing qualitative data, founded in part on discourse analysis, and we address:

- discourse analysis and critical discourse analysis
- conversational analysis
- narrative discourse
- autobiography

These approaches do not use coding and they keep together the text rather than fragmenting it as in coding. In each instance we provide a worked example, so that readers can understand the issues more clearly.

35.1 Discourse analysis and critical discourse analysis

Words carry many meanings; they are nuanced and highly context-sensitive. Meanings are always contextual; what we say, hear, write and read are always connected to their contexts (Andrelchik, 2016, p. 137), for example, physical, temporal, interpersonal (shared understandings and evaluations), cultural, societal, valuative, ideological etc., and to the hidden assumptions in which they are embedded. As Denscombe (2014) suggests, we can never take qualitative data, in this case texts, at face value; they have to be deconstructed to expose their hidden messages (p. 288).

In qualitative data analysis, interpretation and analysis are often fused and, indeed, concurrent or simultaneous. It is naive to suppose that the qualitative data analyst can separate analysis from interpretation, because words themselves are interpretations and open to interpretation. In this chapter we show how qualitative researchers can analyse discourses, be they in written texts or transcriptions of spoken conversations.

'Discourse' is a slippery term. It designates how language represents meanings, conventions, codes in specific socio-cultural, temporal and historical contexts, how linguistic practices are both located in, and create, their own contexts (Hammersley, 2013). Discourses shape and are shaped by different meanings, and people are members of different discourse communities – those communities which hold similar values, views, ideas and ways of looking at the world. Discourses are 'social texts … particular signifying practices of a given group [that] are both constituted by and constitutive of the discursive field in which members of the group live and function' (Elbaz, 1990, p. 15).

Discourse analysis is an umbrella term with many different meanings and types, and we address some of these here.

Discourse analysis can include linguistic matters in their cultural, ideological, political (micro and macro), historical and societal contexts, and in the assumptions in which they are embedded. Denscombe (2014, p. 290) suggests several issues which discourse analysts can address, for example:

- the discourse in context (variously defined as above), be it in the form of text, conversation, narrative, biography etc., and in terms of which groups and conditions they represent;
- how power, interests and influence operate through language;
- whose perspective/version is being portrayed in the discourse, and what alternatives are possible;
- what is absent, silenced, neglected or suppressed in the discourse;
- what linguistic devices are present in the discourse.

We distinguish between discourse analysis and critical discourse analysis.

Discourse analysis

At one level discourse analysis investigates language and the meanings that are given to texts which create and shape knowledge and behaviour (cf. Sinclair and Coulthard, 1975; Lemke, 1989; Gee, 2005). It examines language in use, linguistic features and forms, language patterns and units, and how meanings are constructed through texts beyond the single sentence level (cf. Wetherell *et al*., 2001; Souto-Manning, 2014).

Any text can be the bearer of several discourses, and a single text can be deconstructed into several meanings. Texts are set in social contexts (Gee, 1996, 2005) and reality is a social construction, so discourse analysis has to take account of the social contexts in which the texts are set. Hence, at another level, discourse analysis looks beyond linguistic features to the links between language and society, language and the social context in which they are set. We give examples of these below, in conversational analysis, narrative analysis and autobiographical analysis. We also introduce critical discourse analysis.

To constitute a discourse, Renkema (2004) suggests that a text – spoken or written – must fulfil seven main criteria:

1. *Cohesion*: there must be a grammatical relationship between the different parts of the text or conversation.
2. *Coherence*: the sequence and structure of the text must make sense and 'hang together'.
3. *Intentionality*: the text or conversation must be written or spoken intentionally.
4. *Acceptability*: it has to be accepted or acceptable to its intended audience.
5. *Informativeness*: it must include new information.
6. *Situationality*: the context, conditions and circumstances in which it is embedded must be known and made explicit.
7. *Intertextuality*: the text or conversation must go beyond simply the text to an outer world of the reader, interpreter, researcher and other agents.

Discourse analysis regards talk and texts as social practices (Potter and Wetherell, 1994, p. 48), agentic, interactive and socially constructivist (Clifton, 2006). Discourse analysis is influenced by speech act theory (Austin, 1962; Searle, 1969) of locutions (what is uttered), illocutions (doing something whilst saying something) and perlocutions (achieving something by saying something), and by textual analysis, ideological analysis and ideology critique (Potter and Wetherell, 1994, p. 47). Here language is not simply a 'representation of an inner mental state or a conduit for one's thoughts' (Paulus and Lester, 2016, p. 408); rather it is used to do something, to achieve something; it is oriented to action (p. 408). A discourse is language in use as a social practice (Fairclough, 1992).

The researcher analysing discourses in their linguistic form can examine, for example: meanings, form, style genre, register, order, cohesion, episodes, metaphors, categories, aporias, metonyms, key issues, repetitions, mimesis, shibboleths and networks (e.g. Lemke, 1989; Tunnicliffe and Reiss, 1999; Andrelchik, 2016).

Wetherell *et al*. (2001) add to this the broader social context, identifying four methods of discourse analysis, analysing:

1. words in context (e.g. cultural, social, group) as ways in which people express themselves and in which context influences the language used, i.e. how context affects meaning and language;
2. interactions conducted through language;
3. patterns of language use (e.g. language used to express wishes, emotions, reactions, to create scenarios, to give information);
4. links between language and the constitution, structure and nature of society, often focusing on differentials of power and their reproduction.

Texts themselves carry many levels of meaning, and the qualitative researcher must strive to catch these different levels or layers. Further, researchers are often part of the world that they are actually writing about, and, even if they are not, they bring their own culture, norms and values to bear in conducting, analysing, interpreting and reporting the research. Issues of projection and countertransference are important: the researcher's analysis may say as much about the researcher as about the text being analysed, both in the selection of the levels of analysis, the actual analysis, and the inference of intention and function of discourses in the text, with their corollary in the key issue of reflexivity.

Critical discourse analysis

Critical discourse analysis moves beyond the purely linguistic level set out above, and examines the exercise of power through discourses, texts, conversations and narratives (e.g. Fairclough, 1995; Gee, 1996), i.e. how power operates through all types of discourses. Here language is not neutral and 'there are no "neutral" words and forms' (Bakhtin, 1981, p. 293). Hence critical discourse analysis links language and society, and is a way of thinking, perhaps culturally or institutionally conditioned, which, like a paradigm, is legitimated by communities, often those with power. Critical discourse analysis reveals how power operates and is constituted, shaped, legitimated, maintained, regulated and challenged in and through language and discourses (e.g. Fairclough, 1992, 1995; Fraser, 2004). As Foucault (1998, p. 101) remarks, 'discourse can be both an instrument and an effect of power'; it is the 'tactical dimension' of the operation of power in individuals, groups and organizations.

Power is immanent in discourse; it is one of its defining features, and power relations are intrinsically discursive. Indeed the three examples in this chapter all

concern power, its possession, denial, operations, fluidity, negotiation, relations, absence, and so on. In critical discourse analysis, researchers examine how language 'colonized everyday life through institutional discourses' (Souto-Manning, 2014, p. 160), i.e. how power, operating through language, reproduces power differentials in society and the lifeworlds of its members (the 'taken-for-granted universe of daily social activity' which 'always remains in the background') (Habermas, 1987b, p. 131).

Critical discourse analysis, stemming in large part from the Frankfurt School of critical theory (see Chapter 3), is linked to ideology critique of power and power relations, interests and their operations, and has an explicit agenda of critiquing inequalities, discrimination and ideological domination; it seeks to transform and emancipate society and its members, and redress illegitimate imbalances of power and influence within relationships. It interrogates ideological, political, social and economic power and how this is created, achieved, perpetuated and reproduced through discourses.

Critical discourse analysis works with, for example, the voices of marginalized, disempowered and oppressed groups, and it critiques the illegitimate power of dominant groups and the role of language in this. We discuss critical theory much more fully in Chapter 3, and we advise readers to review that chapter. Critical discourse analysis looks at a social problem, not just a research question (Fairclough, 2003), and uses linguistic analysis to identify and expose ideology and power at work in society. It links micro- and macro-analysis, and this differentiates it from purely linguistic discourse analysis which typically operates much more at the micro-level.

In critical discourse analysis, texts are interrogated not only for what they include but what they exclude: structured silences and how these embody differentials of power and influence in society. They also concern what is implied, though not overtly spoken. In other words, the researcher must read along, between and beyond the lines.

Below we take three examples of ways in which researchers can conduct discourse analysis: a conversation, a narrative and an autobiographical text. A conversation involves more than one person; a narrative is written by a single person; and an autobiography is a narrative that is written by, and in, the first person.

35.2 A conversational analysis

Conversation analysis is one type of discourse analysis, looking at a conversation between two or more people in a specific context, examining what they say, how they say it, for what reasons or purposes, and using what kinds of interaction, sequences, contexts and structures in the conversation. It is the 'formal analysis of everyday situations' (Flick, 2009, p. 334), how participants create meanings of their conversational situations and achieve their intended actions and outcomes, and how they get things done through conversations (e.g. speech acts). Though some forms of conversational analysis have been criticized for their focus on the trivial (Hammersley, 2013), there is a much richer story to be told, as we indicate below.

Conversation analysis can examine turn-taking, sequences and the evolution of a conversation, interaction in conversation, cohesion, purposes of conversations and the expectations of participants, language rights and roles in conversations, strategizing in and through language. Further, it can focus on issues of power, domination and the constructions and reproduction of power in texts and conversations, and language in social contexts and interactions.

Conversational analysis is a rigorous investigation of features of a conversation, how it is generated and constructed, how it operates, what its distinguishing features are, how participants construct their own meanings in the conversational situation (Clifton, 2006, p. 203), and how conversations are located within their several contexts (Vaughan, 2012). The following example of a conversational analysis exposes the multi-leveled interpretations that can be made of conversations as discourses.

The example is of a transcript of a short conversation in an infant classroom (Cummings, 1985) which contains the potential for several levels of analysis; several meanings can be deconstructed from this conversation, and some of them concern power. The analysis also raises the issue of reflexivity in the researcher. The example is in the tradition of Sinclair's and Coulthard's (1975) seminal work on 'institutional talk' in classrooms. Here, as in many examples of conversational analysis, a single episode is analysed.

This is a class of twenty-seven 5–6-year-old children, with the children seated on a carpet and the teacher seated on a chair. A new set of class books has arrived for the children's free use. After a few days the teacher feels that the class and the teacher should look at them together.

Let us explore the levels of analysis here. If we ask 'what is being learned here by the children?' there are

BOX 35.1 TRANSCRIPT OF A CONVERSATION IN AN INFANT CLASSROOM

1	T	Right. Let's have a look at this book – 'cause these are – smashing books. Are you enjoying them?
2	CC	Yes // Yes // Yes.
3	T	What's it called this one? Can anyone tell me?
4	CC	Splosh//.
5	C	//Splish//
6	CC	//Splosh//
7	T	Splosh not splish. It's got an 'o' in the middle. Splosh.
8	CC	Splish splosh//
9	C	//Splosh//
10	T	Splosh it says. (Reading) A dog, a pig, a cow, a bear, a monkey, a donkey, all in the –
11	T & CC	Air
12	T	((Showing pictures)) There's the dog and the pig and the cow and the bear and the monkey and the donkey all in the air. What are they in the air in?
13	CC	()//
14	T	//Put up your hand if you know. Vicky. ((Buzz of children trying to get in))
15	C	The cow's popped it
16	Vicky	// A hot air balloon.
17	T	A hot air balloon
18	C (as 15)	The cow's popped it.
19	T	What's the cow popped it with?
20	CC	Horn//horn//ear//horn//his horn.
21	T	His horn – it's not his *ear* is it – his ears//
22	CC	((Laughing))//
23	T	are down here. It's his horn that's sticking up.
24	CC	((Laughing))
25	T	What does this mean then? ((showing stylized drawings of air escaping))
26	C	Air's coming out//
27	C	//Air//
28	T	The air coming out of the balloon isn't it. Can you *really* see the air coming out of a balloon?
29	CC	No. No. No.
30	T	No – very often in cartoons it look like that doesn't it.
31	C	I can see gas coming out of my mouth when I () on the windows.
32	T	*When* can you see it?
33	C	When it's steamed up.
34	T	Yes. And if//

continued

continued

35	C	//When it's cold.
36	T	When it's cold. When you hhh//
37	C	//When your breath – when your breath turns over and it steams on the – steams on the window.
38	T	Yes//
39	C	And it//
40	T	But only when it's –
41	CC	Cold.
42	T	Cold. Only when it's cold.
43	C	I saw a airship.
44	T	Did you. When? Where?
45	C	On the park.
46	T	Really.
47	CC	I have // I saw // Mrs. Cummings
48	T	Shh – Yes, Luke.
49	Luke	When we – when the airship was aft – when it was finished and the Pope was on we took the telly outside – and – we took the telly outside – and – and we saw – we saw the good old airship.
50	T	Did you.
51	Luke	An air balloon as well.
52	T	It's not good *old* airship – it's Goodyear – the Goodyear airship.
53	CC	Good year // Mrs. Cummmings
54	T	Good year. Yes.
55	C	I seed the airship. ((Many children talking at once))
56	T	Just a moment because I can't hear Luke because other people are chattering. You'll have your turn in a minute.
57	Luke	I said Mummy, what's that thing with the 'X' on the back and she didn't answer me but when I () it off () an air balloon.
58	T	Yes. It was an airship. Yes. Actually I think we saw it at school one day last summer, didn't we.
59	CC	Yes.
60	T	We all went outside and had a look at it. It was going through the sky.
61	CC	()//
62	Luke	Mrs Cummings //
63	C	()
64	T	Uuhm – Ben
65	Ben	I remember that time when it came () over the school.
66	T	Did you. Y-//
67	Ben	//() the same one came over my house when I went home.
68	T	Yes. Paul.

69	Paul	I went to a airship where they did //
70	Luke	//It flew over my house ()//
71	T	//Just a moment Paul because Luke is now interrupting. We listened to him very carefully. Now it's his turn to listen to us.
72	Paul	I went to see a airship where they take off and when I – when I got there I saw () going around.
73	T	Oh … What keeps an airship up in the air?
74	CC	Air//air//gas//
75	Luke	Mrs Cummings.
76	T	Air or gas. Yes. If it's air, it's got to be *hot* air to keep it up – or gas. Now put your hands down for a minute and we'll have a look at the rest of the book. ((Reading)) Help said Pig. There he is saying help. ((There is a cartoon-like 'bubble' from his mouth with 'help' written in)) Help said –
77	CC	Monkey
78	T	Help said donkey. It's gone wonky.
79	CC	h-h-h ((untranscribable talk from several children))
80	T	Look as though it had gone wonky once before. What makes me say that?
81	C	Because – because there's – something on the balloon.
82	T	Mmm. There's already a patch on it isn't there to cover a hole ((reading)) A bear, a cow, a pig, a dog, a donkey and a monkey *all – in – a* – and this is the word you got wrong before – all in a –
83	C	Bog
84	T	Bog – Who said it said dog at the end and it shouldn't?
85	James	Me.
86	T	*James*! James, what does it start with?
87	James	'b' for 'bog.
88	T	'b'. It only goes to show how important it is to get them the right way round//
89	C	//Toilet//
90	T	No. I don't think it means toilet.
91	CC	((Laughter))
92	T	I don't think they're in a toilet.
93	CC	((Laughter))
94	T	What's a bog when it isn't a toilet?
95	Gavin	My brother call it the bog.
96	T	Yes. Lots of people do – call a toilet a bog but I don't think that's what this means.
97	Paul	(fall in) something when – when it sticks to you.
98	T	Yes, you're quite right Paul. It's somewhere that's very sticky. If you fall in its very sticky //
99	C	()
100	T	It's not glue

continued

continued

101	C	It's called a swamp.
102	T	Swamp is another word for it, good boy – but it's not glue, it's usually mud or somewhere. It's usually somewhere – somewhere in the countryside that's very wet. ((Many children talking))
103	C	Mrs. Cummings what ()
104	T	Just a moment you are forgetting to listen. You *are* remembering to think and to talk but you're forgetting to listen and take your turn. Now Olga.
105	Olga	Once my daddy –

Source: Cummings (1985)

several kinds of response. At a formal level, first, there is a *curricular* response: the children are learning a little bit of language (reading, speaking, listening, vocabulary, spelling, letter orientation (e.g. 'bog' and 'dog')), science (condensation, hot and cold, hot air rising, hot air and gas-filled balloons) and soil (a muddy swamp). That concerns the academic curriculum, as it were.

At a second level the children are learning other aspects of development, not just academic but personal, social, emotional and interpersonal, for example turn-taking and organization, exchanging, asking questions, giving responses, cooperation, shared enjoyment, listening to each other, contributing to a collective activity, taking risks with language (the risqué joke about the word 'bog' with its *double entendre* of a swamp and an impolite term for a toilet), and action formation (doing and achieving things through language).

At a third level one can see language rights, nature and uses in the classroom (cf. Edwards, 1980; Walsh, 2006; Vaughan, 2012). Here the text usefully provides numbered lines to assist analysis and to preserve the chronology of events, an essential feature of conversational analysis. One can observe the following, using a closer textual analysis:

- A great deal of the conversation follows the sequence of teacher → student → teacher → student and so on (e.g. lines 28–48); here the analysis of the sequence of the conversation is important. Here it is rare for the sequence to be broken, for instance teacher → student → student (e.g. lines 3–7 and 14–16), and where the sequence is broken, it is at the teacher's behest, and with individual children only (lines 48–52, 64–9, 84–8, 94–8). Where the conventional sequence is broken without the teacher's blessing the teacher intervenes to restore the sequence or to control the proceedings (lines 54–6, 70–1, 103–4).
- It appears that many of the twenty-seven children are not joining in very much – the teacher only talks directly to, or encourages to talk, a few named children individually: Vicky, Luke, Ben, Paul, James and Olga.
- There are almost no instances of children *initiating* conversations (e.g. lines 43, 65, 101); most of the conversations are in *response* to the teacher's initiation (e.g. lines 3, 11, 20, 25, 28, 32, 34, 36 etc.); again the analysis of the sequence of the conversation is important here.
- The teacher only follows up on a child's initiation when it suits her purposes (lines 43–6).
- The teacher teaches the children about turn-taking (lines 56, 71).
- Nearly everything goes through, or comes from the teacher who mediates everything.
- Where a child says something that the teacher likes or is in the teacher's agenda for the lesson, then that child is praised (e.g. lines 34, 42, 54, 58, 76 and 96, 98 (the word 'yes'), 102) and the teacher repeats the child's correct answer (e.g. lines 16–17, 20–1, 29–30, 35–6, 41–2).
- The teacher feeds the children with clues as to the expected answer (lines 10–11, 40–1, 76–7, 82–3).
- Where the conversation risks being out of the teacher's control the teacher becomes much more explicit in the classroom rules (e.g. lines 56, 71, 104); again the sequence of the conversation is important here.
- When the teacher decides that it is time to move on to get through her agenda she closes off further discussion and moves on (line 76); as before, the analysis of the sequence of the conversation is important here.
- The teacher is prepared to share a joke (lines 90–3) to maintain a good relationship but then moves the conversation on (line 94).

- Most of the conversation, in speech act terms, is perlocutionary (achieving the teacher's intended aim of the lesson) rather than illocutionary (an open-ended and free-range, multi-directional discussion where the outcome is unpredictable).
- The teacher talks a lot more than the children.

At a fourth level, employing speech act theory, we can see how some utterances in the conversation are intended not only to involve the children but, thereby, to control them. Lines 76, 82 and 102 show the teacher taking charge of the conversation by talking a lot, which has the effect of keeping the children quiet and of reining in the children's talk: a perlocutionary speech act that reasserts classroom control through talk, and it is noticeable that this is later in the conversation rather than earlier, as the children may be starting to become restless. Then, in line 104, when that strategy has not worked particularly effectively, the teacher takes a more overt control strategy and tells children to listen and take turns.

At a fifth level, one can begin to theorize from the materials here. It could be argued, for example, that the text discloses the overt and covert operations of power, to suggest, in fact, that what the children are learning very effectively is the hidden curriculum in which power is a major feature, for instance:

- The teacher has the power to decide who will talk, when they will talk, what they will talk about and how well they have talked (cf. Edwards, 1980).
- The teacher has the power to control many children (twenty-seven children sitting on the floor whilst she, the teacher, sits on a chair, i.e. physically above them).
- The teacher controls and disciplines *through* her control of the conversation and its flow, and, when this does not work (e.g. lines 56, 71, 104) then her control and power become more overt and naked. In this sense it is important to note that conversational analysis often addresses the sequence of the conversation, and this is pertinent here: once a gentle control strategy does not work a more overt strategy is brought into play. What we have here is also an example of Bernstein's (1975) 'invisible pedagogy', for example: where the control of the teacher over the child is implicit rather than explicit; where the teacher arranges the *context*; and where there is a reduced emphasis upon the transmission and acquisition of specific skills.
- What we have here is an example of how talk in classrooms is 'institutionalized' (Sinclair and Coulthard, 1975), and of the importance of the children learning the hidden curriculum of classrooms (Jackson, 1968), wherein they have to learn how to cope with power and authority, praise, denial, delay, membership of a crowd, loss of individuality, rules, routines and socially acceptable behaviour. As Jackson says, if children are to do well in school then it is equally, if not more important that they learn and abide by the hidden curriculum rather than the formal curriculum.
- What we have here is also an example of Giddens's (1976, 1984) structuration theory, wherein the conversation in the classroom is the cause, the medium and the outcome of the perpetuation of the *status quo* of power asymmetries and differentials in the classroom, reinforcing the teacher's control, power and authority.
- The teacher has been placed in a difficult position by being the sole adult with twenty-seven children, and so her behaviour, motivated perhaps benevolently, is, in fact a coping or survival strategy to handle and manage the discipline with large numbers of young and demanding children – crowd control.
- The children are learning to be compliant and that their role is to obey, and that if they are obedient to a given agenda then they will be rewarded.
- The 'core variable' (in terms of grounded theory') is 'power': the teacher is acting to promote and sustain her power. When it can be asserted and reinforced through an invisible pedagogy then it is covert; when this does not work it become overt.

Now, one has to ask whether, at the fourth level, the researcher is reading too much into the text, over-interpreting it, driven by her own personal hang-ups or negative experiences of power and authority, and over-concerned with the issue of discipline, projecting too much of herself onto the data interpretation. Maybe the teacher is simply teaching the children socially acceptable behaviour and moving the conversation on productively, exercising her professional task sensitively and skilfully, building in the children's contributions, and her behaviour has actually nothing to do with power. Further, one can observe at level four that several theories are being promulgated to try to explain the messages in the text, and one has to observe the fertility of a simple piece of transcription to support several grounded or pre-ordinate/pre-existing theories. The difficult question here is, 'which interpretation is correct?' Here there is no single answer; perhaps they are all correct.

The classroom transcription only records what is said. People will deliberately withhold information; some children will give way to more vocal children, and others may be off task. What we have here is only one medium that has been recorded. Even though the transcription tries to note a few other features (e.g. children talking simultaneously), it does not catch all the events in the classroom. How do we know, for example, whether most children are bored, or if some are asleep, or some are fighting, or some are reading another book and so on? All we have here is a selection from what is taking place, and the selection is made on what is transcribable.

One can see in this example that the text is multilayered. At issue here are the levels of analysis that are required, or legitimate, and how analysis is intermingled with interpretation. In qualitative research, analysis and interpretation frequently merge. This raises the issues of validity and reliability. What we have here is the 'double hermeneutic': as researchers we are members of the world that we are researching, so we cannot be neutral; we live in an already-interpreted world. Look at the example above:

- The teacher and the children act on the basis of their interpretations of the situation (their 'definitions of the situation').
- The lived actions are converted from one medium (observations, actions and live events) to another (written) by choosing to opt only for transcription: an interpretation of their interpretation.
- The researcher then interprets the written data (a third hermeneutic) and writes an unavoidably selective account (a fourth – quadruple – hermeneutic – an interpretation of an interpretation of an interpretation of an interpretation!).
- The reader then brings his/her own biography and background to interpret the researcher's written interpretation (a fifth – quintuple – hermeneutic).

Given the successive interpretations, it is difficult not to suggest that reliability and validity can easily be compromised in qualitative research. Reflexivity, as the disclosure of one's possible biased interpretations, does little to reduce them; I can state my possible biases and interpretations but that does not necessarily stop me or them from being selective and biased. This suggests, perhaps, the limits of reflexivity. In connection with increasing reliability and validity, reflexivity is not enough.

35.3 Narrative analysis

Narrative analysis, as with discourse analysis, encompasses different approaches which adopt differing ontological and epistemological positions on the social world and how it is construed and constructed (Gee, 2005; Kennedy-Lewis *et al*., 2016). A narrative is a story with an individual perspective, written in the teller's own voice, in which the teller controls what is released, when and in what sequence. A narrative can also process and condense large amounts of data, to provide a 'more complex and complete picture of social life' (Hendry, 2007, p. 489); it has a purpose, a plot and a human element (Denscombe, 2014, p. 291). Narrative analysis creates a unity out of disparate elements; it creates a story. Narratives often include and evoke emotional and aesthetic elements and responses respectively (Rogan and de Kock, 2005; Barone, 2007; Hendry, 2007; Rosiek and Atkinson, 2007). As Rosiek and Atkinson write: 'humans generally live a storied existence' (p. 503). Indeed Hendry (2007) suggests that narratives are 'highly seductive' (p. 488) and a form of democratic research (p. 490).

We distinguish here between: (a) 'narrative analysis' (Polkinghorne, 1995, p. 12), where the researcher produces a narrative from data, events, happenings and information from various sources to create a plot, a thematic line a temporal sequence or structure; and (b) 'analysis of narratives' (p. 12), where the researcher analyses the narratives produced by other parties. In the former, the narrative is the *consequence* of the research whilst in the latter it is the *source* of the researcher's knowledge (Smeyers and Verhesschen, 2001, p. 76). The example below is of the latter. However, the comments we make here apply to both (a) and (b).

Polkinghorne (1995) describes a narrative as 'a type of discourse composition that draws together diverse elements, happenings, and actions of human lives into thematically unified goal-directed processes' (p. 5). Connelly and Clandinin (1999) broaden Polkinghorne's definition to move beyond simply stories and to include the research process and the researchers' interpretation of the story of the research itself, though Smeyers and Verhesschen (2001) caution against adopting too wide a definition, as it would mean that anything could count as a narrative (p. 78).

A narrative analysis reports personal experiences or observations and brings fresh insights to often familiar situations; narrative text has an 'omniscient, authorial voice' (Bruner, 2004, p. 702). It is strongly interpretivist, the author's construction rather than an objective truth (Smeyers and Verhesschen, 2001), with meanings constructed through observations and

language. Indeed it is sometimes difficult to separate facts from observations, as many narratives can use data selectively and report them in non-neutral terms (as in the example that follows). As with other forms of discourse analysis, narrative analysis is rooted in a social constructivist paradigm in which behaviours and their meanings are socially situated and socially interpreted.

Riessman (2008) suggests that narrative analysis can use *thematic analysis* (identifying categories and themes), *structural analysis* (how the narrative is structured and what the language does at textual and cultural levels), *performance analysis* (how narratives are co-constructed/done/performed and the difficulties encountered in such structuring) and *visual analysis* (of narratives constructed using visual media).

The example that follows is taken from Goffman's (1968) *Asylums* (a study of a psychiatric hospital). Here the 'asylums' – hospitals – bear many similarities to schools, particularly boarding schools, in being 'total institutions'. By taking a non-school example here, it is intended to 'make the familiar strange' (Blumer, 1969): to make the familiar world of schools 'strange' to the researcher (i.e. to see schools with a new eye) by comparing them to another similar but also different institution.

Goffman (1968, pp. 17–19) writes that a total institution (e.g. a hospital, an army, a boarding school, a prison), is characterized by several features:

- The institution is convened for a specific purpose.
- All aspects of life take place in the same place and under the same single authority.
- Every part of the member's normal daily activities takes place in the company of many others.
- All members are treated the same and are required to do the same things together.
- The daily activities are precisely and tightly scheduled by a controlling authority and officials, and through formal rules that are tightly enforced.
- The several activities are part of a single, overall plan that is intended to fulfil the aims of the organization.
- There is a division between the managers and the managed (e.g. the inmates and the hospital staff; the teachers and the students).
- The inmates have limited or no contact with the outside world but the officials do have contact with the outside world.
- Access to the outside world for inmates may be physically or institutionally restricted, controlled or forbidden.
- There is some antagonism between the two groups, who hold hostile stereotypes of each other and act on the basis of those stereotypes, often based on inequalities of power.
- Officials tend to feel superior and powerful whilst inmates tend to feel inferior and powerless.
- The cultures and cultural worlds of the officials and the inmates are separate.
- The two worlds – of officials and inmates – have limited penetration of each other.
- There is a considerable social distance between the two groups.
- Inmates tend to be excluded from knowledge of decisions made about them.
- Incentives (for work, behaviour) and privileges have greater significance within the institution than they would in the outside world.
- There are limited and formal channels of communication between the members of the two worlds.
- Release from the institution is often part of the privilege system.

These features can apply to several different total institutions, of which schools are an example.

Goffman (1968, pp. 220–5) presents a narrative account of his field notes on the psychiatric hospital, synthesized into a single text.

> In everyday life, legitimate possessions employed in primary adjustments are typically stored, when not in use, in special places of safekeeping which can be gotten to at will, such as foot-lockers, cabinets, bureau drawers, and safe-deposit boxes. These storage places protect the object from damage, misuse, and misappropriation, and allow the user to conceal what he possesses from others …
>
> When patients entered Central Hospital, especially if they were excited or depressed on admission, they were denied a private, accessible place to store things. Their personal clothing, for example, might be stored in a room that was beyond their discretionary use. Their money was kept in the administration building, unobtainable without medical and/or their legal agents' permission. Valuables or breakables, such as false teeth, eyeglasses, wrist watches, often an integral part of body image, might be locked up safely out of their owners' reach. Official papers of self-identification might also be retained by the institution. Cosmetics, needed to present oneself properly to others, were collectivized, being made accessible to patients only at certain times. On convalescent wards, bed boxes were available, but since they were unlocked they were subject to theft from other patients and from staff, and in any case were often located in rooms locked to patients during the day.

If people were selfless, or were required to be selfless, there would of course be a logic to having no private storage places, as a British ex-mental patient suggests:

> I looked for a locker, but without success. There appeared to be none in this hospital; the reason soon [became] abundantly clear; they were quite unnecessary – we had nothing to keep in them – everything being shared, even the solitary face cloth which was used for a number of other purposes, a subject on which my feelings became very strong.

But all have some self. Given the curtailment implied by loss of places of safekeeping, it is understandable that patients in Central Hospital developed places of their own.

It seemed characteristic of hospital life that the most common form of stash was one that could be carried around on one's person wherever one went. One such device for female patients was a large handbag; a parallel technique for a man was a jacket with commodious pockets, worn even in the hottest weather. While these containers are quite usual ones in the wider community, there was a special burden placed upon them in the hospital: books, writing materials, washcloths, fruit, small valuables, scarves, playing cards, soap, shaving equipment (on the part of men), containers of salt, pepper, and sugar, bottles of milk – these were some of the objects sometimes carried in this manner. So common was this practice that one of the most reliable symbols of patient status in the hospital was bulging pockets. Another portable storage device was a shopping bag lined with another shopping bag. (When partly full, this frequently employed stash also served as a cushion and back rest.) Among men, a small stash was sometimes created out of a long sock: by knotting the open end and twisting this end around a belt, the patient could let a kind of moneybag inconspicuously hang down inside his trouser leg. Individual variations of these portable containers were also found. One young engineering graduate fashioned a purse out of discarded oilcloth, the purse being stitched in separate, well-measured compartments for comb, toothbrush, cards, writing paper, pencil, soap, small face cloth, toilet paper – the whole attached by a concealed clip to the underside of his belt. The same patient had also sewn an extra pocket on the inside of his jacket to carry a book. Another male patient, an avid newspaper reader, invariably wore a suit jacket, apparently to conceal his newspapers, which he carried folded over his belt. Still another made effective use of a cleaned-out tobacco pouch for transporting food; whole fruit, unpeeled, could easily be put in one's pocket to be taken back to the ward from the cafeteria, but cooked meat was better being carried in a grease-proof stash.

I would like to repeat that there were some good reasons for these bulky carryings-on. Many of the amenities of life, such as soap, toilet paper, or cards, which are ordinarily available in many depots of comfort in civil society, are thus not available to patients, so that the day's needs had to be partly provided for at the beginning of the day.

Fixed stashes, as well as portable ones, were employed, too; they were most often found in free places and territories. Some patients attempted to keep their valuables under their mattresses but, as previously suggested, the general hospital rule making dormitories off-limits during the day reduced the usefulness of this device. The half-concealed lips of window sills were sometimes used. Patients with private rooms and friendly relations with the attendant used their rooms as stashes. Female patients sometimes hid matches and cigarettes in the compacts they left in their rooms. And a favourite exemplary tale in the hospital was of an old man who was claimed to have hidden his money, $1,200 in a cigar box in a tree on the hospital grounds.

It would be plain that some assignments also provided stashes. Some of the patients who worked in the laundry availed themselves of the individual lockers officially allocated only to non-patient workers. The patients who worked in the kitchen of the recreation building used the cupboards and the refrigerator as places in which to lock up the food and drink they saved from the various socials, and other indulgences they had managed to acquire.

(Goffman, 1968, pp. 220–5)

The narrative account tells a gripping, disturbing story in much more graphic detail than would be possible through the often decontextualized world of extracted, codified and reassembled data; the narrative makes the most of the virtues of a story: an account that 'catches fire' through the language used, that persuades, that is human, that is rich in detail and that tells a story. What is that story?

At first sight the patients' behaviour may seem very odd, they seem fixated on minute matters, they dress bizarrely, their clothing bulges with a range of objects that 'normal' people would not carry around, they are

obsessive about hoarding, they trust nobody, and what they take so many pains to carry around is almost worthless. They might be rightly accused of not being in their right mind, and therefore that they are rightly incarcerated in the secure hospital so that they are no danger to themselves and to others. That is one version, one discourse.

However, an alternative explanation can be offered, an alternative discourse is at work. Here one can see why their behaviour is as it is. For example, if we look at the descriptions of what the patients were experiencing we can observe:

- their personal clothing was available only at the discretion of the staff;
- valuables were kept locked away from the patients;
- self-identification papers were held by the institution;
- cosmetics were made available only at certain times;
- everyday amenities of life in civil society were not available to the patients;
- bed boxes were kept unlocked, i.e. nothing was secure;
- everything was shared;
- there were no free places;
- private spaces (dormitories) were off-limits during the day;
- individual lockers were for non-patients.

What we see in both an actual and metaphorical sense is the stripping away of identity, personality, individuality, privacy, security, power, freedom, autonomy, humanity and decision making, and all by those with power over the inmates. Goffman (1968) terms this the processes of *depersonalization* and *mortification*. Nothing personal is left to the patients; nothing is private, nothing is safe.

If we look at the vocabulary that Goffman uses in connection with the patients, we see very many terms about these same points: 'stash', 'possessions', 'protect', 'conceal', 'stored', 'storage', 'safekeeping', 'valuables', 'hidden', 'hid', 'half-concealed', 'lock up', 'containers', 'saved'. They have actual and metaphorical meaning: at both an actual and metaphorical level the patients are trying to retain their lost personalities, identities, rights, autonomy and freedoms, even their sanity. It is little wonder, then, that metaphors of storage, protection, privacy, keeping things safe and containment are realized in practice. Indeed it could be argued that, far from being disturbed or out of their minds, the patients were behaving very sanely and sensibly in an insane or disturbing situation. How often do we find the same situation in schools, where students behave very sensibly in the face of extreme or unacceptable behaviour by teachers (but often the blame is placed on 'disruptive' students who dare to disrupt the power-and-control oriented, boring and dominatory behaviour of teachers)? Sanity and madness are, to some degree perhaps, a social construct rather than an objective reality.

Descriptive data in this narrative form enable the researcher to understand the situation vividly from the perspective of the participants, their 'definition of the situation'. The hospital staff might have put a very different interpretation on their own behaviour, arguing that they were removing sources of distress and danger from the patients, and caring for them very extensively. That may be true also; reality is multifaceted. Through an analysis of the narrative, the descriptive data help the researcher to explain why situations are the way that they are. In fact one could argue that the patients are behaving very rationally and reasonably in an unreasonable, power-stripping and depersonalizing situation, even though their behaviour at first might seem strange.

The extract is powerfully written; the structured silence on the less antagonistic or depersonalizing behaviour of the staff and the regime is presented selectively, if at all, but the force of the narrative is the stronger for this. The well-chosen examples of the hiding of even everyday objects are given extraordinary semiological, symbolic power in indicating how power reaches right to the heart of commonplace, almost taken-for-granted matters. The narrative is a well-worked example of how the taken-for-granted, everyday world and its artefacts can have extraordinary meaning in certain contexts. When these everyday objects are used to make grotesque shapes in the clothing of the patients, rendering them instantly recognizable as patients by their freakish garb, the contrastive power of this juxtaposition is startling. Whilst this is not the place to go into semiotics, narrative analysis can use semiotic analysis: the interpretation of signs and symbols as signifiers of meaning.

In examining the narrative, the researcher can look for what is happening, what are the main features being reported, why the behaviours were as they were (and on what basis of evidence the researcher is making that judgement), what other inferences and explanations might be made of the data provided, and what other data might be needed to support or refute the inferences and explanations given.

35.4 Autobiography

Bruner (2004) argues that we regard 'lived time' as a narrative (p. 692), a story that has meaning for us and which shapes our lives (as he remarks (p. 694): 'we become the autobiographical narratives by which we "tell about" our lives'); our own stories direct our future lives (p. 708), they instruct, shape, reveal and inform our lives (Eisner, 1997, p. 6). Or, as Sartre (1964, p. 39) notes: we are surrounded by our own and others' stories and we interpret our lives in terms of these and, indeed, try to live our lives as if we are recounting those stories. Indeed Plummer (1995; 2001) argues that an essential feature of being human is our creation of stories to ourselves and others, and that these are essential features of research inquiry.

An autobiography is, as Bruner (2004) writes, 'a privileged but troubled narrative because it is both subjective and objective, reflective and reflexive', in which the narrator is also the central figure (p. 693). Given this, an autobiographical narrative is multilayered and selective, and it can be deconstructed at many levels: personal, cultural, interpersonal, ideological, linguistic and so on. It has facts, themes, actors, a sequence, a plot, agency, coherence, situatedness and a sense of audience, all of which are elements of a true discourse as set out at the start of this chapter. It is not coolly objective but often a vivid, evocative account (Tedder, 2012).

In the example that follows, the fictitious autobiography tells a personal story in a highly selective and authentic way. Imagine that the researcher had asked the teacher to write a brief autobiography of his experiences as a teacher; what we have here is the teacher's own views, and this indicates the significance that the writer gives to the events selected.

> I had always wanted to teach music to secondary school students. I had played in a school band when I was at secondary school, and had taken piano lessons for ten years, and had passed all the grades, and I thought that it would be really good to teach. I thought it would be good for students to be exposed to the great classics, or modern music, and I thought that it would be even better if I could teach them how to read, write and compose music. I thought that this would be particularly interesting for downtown kids who had not had access to such music, so I was keen to work in an inner city school.
>
> I had been working in business for 25 years, ten years with a printing company and then 15 years in a commercial company selling paper products. But I felt dissatisfied with my life, so I decided to do what I had always wanted to do, which was to train to be a secondary school music teacher. So I discussed it with my family and gave up my job to take a teacher training course. I was very keen and worked hard on my studies.
>
> I was very happy when the course began; we were introduced to all sorts of ways in which students could learn to write, read and play music, how they could work in pairs and groups to devise musical compositions, how to read non-standard musical scores, how to use the electronic instruments that had not been around when I was at school, and how to teach students to appreciate all different kinds and genres of music.
>
> I passed my course and went to a downtown school. What a total let down! The students didn't care about music; they saw it as a waste of time and boring. They thought that the music syllabus was old-fashioned, that it did not represent the music that they were interested in. All they wanted to do was to play to the whole class their own latest music releases and albums from the ridiculous groups and so called 'artists' whom they had seen gyrating sexily on the television and the Internet. When I tried to change the activities, so that they were playing musical instruments and composing their own music, they either just made a whole lot of noise with them, and the din was awful, or they thought that the instruments and the activity were just babyish, so they did nothing except fool around in the class. Everything that I had been taught about discipline in my teacher training didn't work. At first I thought that it was that my class control wasn't very good, so I asked my mentor how to improve this, but it did not help – the students just sat and laughed, shouted, or refused to do anything.
>
> So I tried a different approach: I told them that they had to learn several musical 'facts' such as information about the lives of composers and the names of pieces of music of certain composers. In fact this wasn't so much a music lesson as a reading lesson. I told them that I was going to give them a test on this, and that those who didn't score highly enough on the test would be punished. I thought that by making the lesson more like a 'high status' area of the curriculum, and coupled this with a test, it would make the students take this more seriously, but it didn't. All they said was that they didn't care, that music was a waste of time, and that it wouldn't help them to get a job. I felt very frustrated.
>
> It didn't get better and I was worn out, stressed, and felt that I was in a job that was completely unrewarding. So, in the end I looked for a job in a

private secondary all-boys boarding school, thinking that at least the students would be more motivated and well behaved. I was hopeful and felt good. I got a job in a small private secondary school where the students had to learn a music instrument at school, as well as taking class music lessons. I hoped that this would be the answer, and that I would be happy again and able to teach 'real' music.

However, I quickly found out that this wasn't the solution. Whilst some of the students were motivated and very nice indeed, some of them were arrogant and treated me as a hired servant whom they could control by threatening to report me to the Senior Teacher if I raised my voice to them or set them too much work to do. I felt insulted.

I didn't like their attitude to me or to the subject; I had been told to follow a more traditional curriculum, and I was very happy to do this, but I found that the students thought that the music lessons were 'beneath their dignity', trivial, and 'tame' compared to the other subjects on the curriculum. In turn, at first I thought that they were just young, fashionable, upper-class or would-be upper-class students with a superior attitude and self-confident manner, wealthy, privately educated, privileged, brash, indulgent, with an expensive lifestyle and high living, a love of country sports, and even a shared way of speaking, and I humoured them, but, the longer it went on the more it irritated me, as I felt that they were looking down on me and on the music lessons.

In fact they weren't all like that, and some of them were from poor, middle-class and working-class homes, whose parents wanted to give their children the chances that had not been available to them, and some students had been thrown out of other schools and had been put into this school by anxious and overwrought parents or by parents who were at their wits end in trying to cope with their badly behaved child. These students continued to be badly behaved, but I was told to 'put up with it', as they brought in a lot of money to the school.

I couldn't take it. One day I exploded with them. I insulted them very strongly, called them all upper-class idiots, called the others 'layabouts', shouted that they should treat teachers with a shred of decency, and basically 'lost it'. The class laughed loud and long; they had won. I left the class and walked out of the job.

I feel very dispirited and let down. I feel as though I have a lot to offer to teaching, but there's no way I can offer it under the present system, so I'm getting out. I'm going back to find another job in business and maybe I'll do some part-time music tuition and give piano lessons in the evenings, with motivated kids and in a situation that is under my control, and where my students will learn something other than how to behave badly.

The autobiography has several *themes* (and themes or *leitmotivs* are a feature of narratives): optimism turning to resentment turning to disillusionment; positive to negative; empowerment turning to disempowerment; dreams turning to dust; power shifts (from the writer to the students); achievement and loss; aspiration turning to deterministic frustration; ignorance turning to knowledge; power turning to loss of control; false expectations to growing realism; and so on. It has a purpose and a plot – a cumulative progression of events over time – and a human, emotional side, i.e. key elements of discourse.

We can observe that the narrative employs a chronological, linear sequence which is interrupted only very occasionally to break off into reflection or comment. The writer has chosen to focus on critical events and decisive moments, all of which are autonomously chosen and life-changing. This is an existential journey in which agentic choice struggles to realize itself as planned and which, in the end, leads to resignation in several senses.

If we examine the text we can observe the overwhelming preponderance of the active rather than the passive voice; here is a writer who is existentially alert. We can note the absence of metaphor, the emphasis on the 'facts' of the events, and a 'no-nonsense' approach to getting on with life (albeit selectively chosen and interpreted) rather than reflections, indeed it is only towards the end of the extract that we can detect a sense of deeper reflection in the writer, i.e. that the writer has learned from experience and the reflection on that experience, and has gained a truer knowledge of the 'real' rather than the perceived or desired situation.

We can note the presence of many stative verbs, phrases and their accompanying adjectives to indicate feelings: 'I thought it would be good'; 'I thought it would be even better'; 'I felt dissatisfied'; 'I was very keen'; 'I felt very frustrated'; 'I was in a job that was completely unrewarding'; 'I was hopeful and felt good'; 'I didn't like'; 'I felt insulted'; 'I felt that they were looking down on me'; 'I feel very dispirited and let down'. Here is a writer who is seeking authenticity, self-realization, emotional fulfilment, who is concerned with feelings.

One interpretation of the text is that it reveals a writer who seeks control and the realization of a personal agenda for happiness; when this is challenged, he finds it hard to come to terms with the situation, to accept it or to accommodate to it. Points of conflict

chart the movement in the text from 'me' to 'them', from the writer's agenda to the student. The word 'I' is used fifty-four times in the extract, whereas the word 'them' occurs only fifteen times. Indeed one can suggest that using the contrast of 'I' and 'them' can denote a perhaps antagonistic stance of the writer, a significant divide between the teacher and the student, a power struggle for control of the agenda. Indeed the word 'them' occurs more frequently whenever things are going wrong for the writer.

We can see a distinctly sympathetic choice of prose, in which the writer's own situation is presented sympathetically and in which the report on the students is almost entirely negative: they are the ones who 'let down' the writer, who 'didn't care' about music, who only wanted to play 'music from the ridiculous groups and so called "artists" whom they had seen gyrating sexily on the television and the Internet', who 'did nothing except fool around', who 'just made a whole lot of noise' and 'just sat and laughed' or who were 'just young, fashionable, upper-class or would-be upper-class students' (note the use of the word 'just' – a negative term here), who didn't take the lesson seriously, and so on. The pejorative tone of the writer – sympathetic to one party and highly unsympathetic to the others – constitutes a very one-sided text. Indeed, as Riessman (1993) remarks, silence – what is not spoken or included, what is left out – is as important as what is said or included (cf. Denscombe, 2014, p. 290). The question is whether this is a problem, as the text is authentic, strong in reality and reveals the intense emotions at play in the situation; it surely catches the 'quality' of the situation so prized by qualitative research.

There are a few tell-tale verbs: the early part of the text includes positive, hopeful verbs such as 'wanted', 'worked hard', 'passed', whereas by the final paragraph we have the dramatic verbs 'getting out', 'going back' (the use of 'back' is perhaps a sign of defeat and a retrograde step).

Do we have sympathy with the writer? Do we think that the writer is a 'control freak' who deserves to come to the kind of self-knowledge that becomes clear by the end of the extract? Did the writer simply receive his just deserts or were the outcomes undeserved and a pity? Did the writer deserve what happened? Has the writer really taken any account of the students? Do we think that the writer has been treated badly by the students? Is the writer weak, strong, too strong, too controlling a person?

This is one reading of the text. But a discourse permits many interpretations. The interpretation above has operated at the level of the personal perspective of the writer, and has suggested that issues of power, control and self-realization feature strongly in the text. An alternative reading is that this is an accurate and authentic account of a horrible situation in which a decent, hard-working and committed person is treated very badly by two groups of distasteful students. Another reading could focus on the quality and contents of teacher training and false aspirations that the teacher training might have led the writer to hold. Another reading might be of the text as an insight into the problems of teaching, for example, indicating that teachers face huge problems of stress, disruptive behaviour and appalling treatment by students, that these constitute a major reason for the flight out of teaching and problems of teacher recruitment and retention, and that there are insufficient support systems for teachers in school. Another reading might be that of social class in education, and the perpetuation of deep-seated class structures through the provision and uptake of different kinds of schooling, curricula and education. We bring our own agenda to the reading and deconstruction of texts. Texts are multilayered.

And who is the writer? Is the writer male, female, young, old, single, in a relationship, living with parents or living alone, able-bodied or disabled, easy-going, temperamental, easily stressed, tolerant, outgoing, introverted, sociable or antisocial, white, non-white, politically left-wing or right-wing, working class or middle class, and with what views on education and music, and so on? We don't know. Perhaps if we had known some of these details then our reading of the text would have been different.

35.5 Conclusion

This chapter has introduced alternatives to coding and the collation of segmented data in qualitative data analysis. It has suggested that the holism of complete texts can constitute discourses, and that variants of discourse analysis have to recognize that discourses and texts are multilayered and open to a range of interpretations and deconstructions. The chapter has given three different examples of these, selected not only for their content but also for their exemplification of three main kinds of discourse: a conversation, a narrative text and an autobiographical extract. Discourse analysis has many meanings, included in which is the recurrent theme of power and its operations (Foucault, 1998; Fraser, 2004). Whilst discourses have the attraction of *emic* research, authenticity and rich language, the researcher has to be mindful not only of the effects of this on the reader, but of the reader's own effects on the text. As Riessman (1993, p. 70) explains, how a person relates

his or her story 'shapes how we can legitimately interpret it'. The chapter has indicated that analysis of narrative, discourse-based data has to attend to the fine-grained details of texts (Potter and Wetherell, 1994, p. 58; Hammersley, 2013, p. 61), together with situating these in the social context and milieu in which they are set (e.g. Clifton, 2006). In combining different narratives, patterns and themes, the researcher can note similarities, commonalities and differences, not only in content, but in terms of tone, style, register, genre, vocabulary, audience, settings, contexts, metaphors and intentions. Given this, there is no single privileged, definitive way of analysing discourse or the meanings that surface from it.

Companion Website

The companion website to the book provides additional material, data files and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Analysing visual media

CHAPTER 36

This chapter introduces researchers to key issues in analysing different kinds of visual image, including still and moving images, and artefacts. It uses tools of analysis that have been introduced in previous chapters, such as content analysis and discourse analysis, and it provides an entrée into the next chapter on grounded theory. With reference to analysing visual data, the chapter introduces:

- content analysis
- discourse analysis
- grounded theory
- interpreting images
- interpreting an image: an example
- analysing moving images

We provide an extended worked example of an analysis of a photograph, to clarify key issues in this kind of analysis.

36.1 Introduction

Chapter 35 introduced discourse analysis. Visual media are a form of discourse. Hence the researcher can use some of the analytical tools that are available to quantitative and qualitative data analysts, for example: content analysis (both numerical and qualitative); discourse analysis and grounded theory. We address these below. Computer software (e.g. NVivo, ATLAS.ti) works with visual data as well as textual data and there are several software packages specifically designed for video data, for example, The Observer XT, Orion, Transana and Video Traces. However, as with other software for qualitative data analysis, software 'does not come with an inbuilt methodology' (Mercer, 2010, p. 10).

Analysing visual data is not straightforward, as images (moving or still) concern meaning making and interpretation. Images are polysemous (Johnson and Black, 2012). Consider, for example, Figure 36.1, an apparently objective, factual photograph of a secondary school senior teacher's office, which has been placed into NVivo, with the researcher's commentary.

What is the researcher to make of this commonplace photograph? It is open to multiple, perhaps contradictory interpretations; there is no simple, single, cold observational analysis. The researcher's commentary reads thus:

> Senior teacher's office. It is small: just enough room for a workstation, one other person, limited shelf space, but little else (note: some materials have to be kept above the top shelf in plastic containers). It has functional office furniture, so the only space to personalise it is by pictures on the notice board and the wall. The location of the smaller chair and the lack of space behind the chair suggest a cramped and confined working space (notice the scrape marks on the wall behind the small chair), though the senior teacher has a more comfortable, and larger chair. Not a room to receive the public, e.g. parents or visitors. No window, lighting comes from fluorescent lighting. A room for working in alone rather than sharing.

Then the researcher moves to some more interpretive comments, and many of these remarks project something of the researcher's own self:

> Given the kind of furniture, the lack of space, no room for superfluous materials, furnishings or ornament to personalise the room, cramped conditions, it is more like a monastic cell for private work and functionality alone: concentrate on the work with no distractions or creature comforts. The occupant could be removed at short notice (so little to remove), or, more positively, is it a retreat for the senior teacher for some peace/silence/privacy/sanctuary? Is the school so short of space that it gives the senior teacher such cramped working conditions? Is it disrespectful? Who would sit in such close – almost physical – proximity to the senior teacher?

Note the different possible interpretations here. Is the room about peace, or privacy, or silence, or sanctuary, or comfort, or solitude, or functionality, or disrespect, or impersonality, or retreat, or invasion of personal

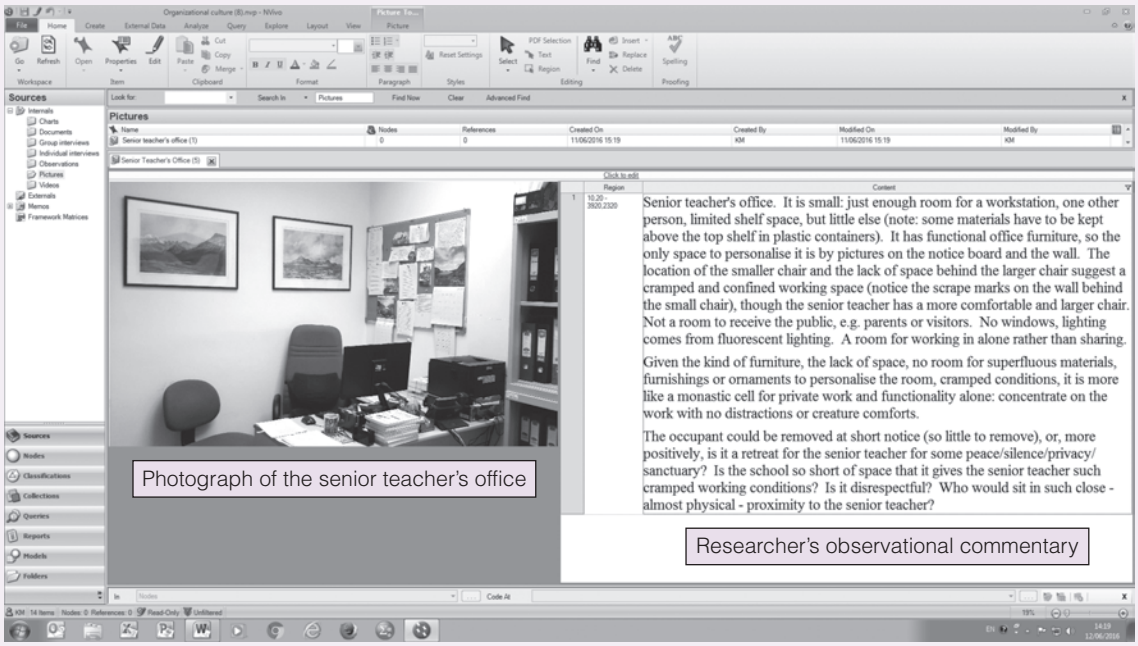


FIGURE 36.1 Picture file for analysing picture data in NVivo (Version 10)

Note

Screenshot reproduced with permission of NVivo qualitative data analysis Software; QSR International Pty Ltd. Version 10, 2012.

space etc.? Is the school trying to sicken the senior teacher by giving her/him a poky little room so that she/he will leave? Do other senior teachers in the school have their own room, and if so, is theirs similar to this? Is the school really so short of space that it is a privilege to have one's own room, regardless of its size and contents?

Maybe the photograph is about all of these, in which case, which interpretation(s) should prevail? The point here is that visual data on their own might be insufficient to draw conclusions, and that it is often wise to use visual data alongside other data in order to support secure conclusions, i.e. to triangulate (see Chapter 31). In analysing visual images, just as with other types of qualitative data analysis, interpretation and analysis run together and it is difficult, if not impossible, to separate them. What starts out as an innocuous image opens floodgates to interpretation. Indeed it is difficult to decide when the process of visual data analysis actually begins (Knoblauch *et al*., 2006), as analysis and interpretation merge into each other.

Before one starts the analytical process, just as with word-based data, it is important for researchers to become very familiar with the data, looking at the images (videos, photographs etc.) several times and immersing themselves in the data, maybe producing some initial narratives about the data. Immersion and its subsequent meaning-making can operate at several levels. For example Kaufmann (2011) suggests four levels of analysis and meaning-making in visual data, in constructing narratives from the images and the 'dynamic interdependence between experience, theory, and power' (p. 8):

- the image itself as interpreted through the technology which produced the image (e.g. video, film, photograph), shutter speed, focus, viewpoint, context of the image;
- the personal, subjective lenses of the researcher;
- the theoretical frameworks used to analyse the image;
- the medium of the message and the meanings that the medium constructs (e.g. video photograph, film etc.).

Denscombe (2014) suggests various strands in analysing visual data:

The image itself: content, genre, styles;
The producer: intentions and context (by whom, when, under what circumstances, why, the intention of the creator);
The viewer: interpretation and context.

(Denscombe, 2014, p. 292)

As he notes, the image can be approached through different lenses: as containing factual information, as a cultural artefact and as a symbolic representation.

With this background in mind we turn to specific ways of analysing visual images.

36.2 Content analysis

We can analyse visual images in a similar way to that of analysing texts, for example, through 'reading' the meanings, through disclosing our own views, perspectives, backgrounds and values (reflexivity). Here content analysis – purportedly an 'objective' form of analysis – can be performed in ways similar to those in qualitative and indeed quantitative data analysis. A possible sequence is set out below:

1. Start with research questions that determine which images (sampling) will be used in the analysis.
2. Retrieve the appropriate images.
3. Decide the unit of analysis. For example, in photographic or still image data it may be the group, the individual, the object; in video data it may be the setting, the time and the events (e.g. Lee *et al*., 2015).
4. Devise a coding system and codes (which must be mutually exclusive, exhaustive and enlightening) (Rose, 2007, p. 65).
5. Code the images according to the codes (cf. Johnson and Black, 2012; Lee *et al*., 2015).
6. Count codes and their frequencies.
7. Reflect on what the coding and the frequencies have indicated.

The researcher can look for patterns in the data, for example over codes, over images, over time, which patterns vary or do not vary, how the patterns are organized etc. (Knoblauch *et al*., 2008).

A celebrated example of this approach is from Lutz and Collins (1993), who examined some 600 visual images in the magazine *National Geographic*. They devised twenty-two predetermined codes to analyse the photographs (e.g. smiling, gender of adults, group size, skin colour, activity, surroundings of people, wealth indicators etc.). From their analysis they concluded that westerners defined non-westerners in terms that made them very different from westerners and 'as everything that the West is not' (Rose, 2007, p. 67) (akin to Edward Said's (1978) notion of the 'other'), as 'natural', less advanced technologically, more attuned to their environment, more spiritual, more exotic and, indeed, naked. The photographs avoided negative imagery (e.g. of poverty, wars, starvation, conflict, illness, physical deformity); in short a sanitized, non-disturbing, non-upsetting and unreal view of non-westerners was portrayed. Issues of power, of dissatisfaction were simply excluded; a structured silence that acted ideologically to reproduce the *status quo* of inequality within and across countries.

Content analysis, as its name suggests, is more concerned with the contents of the image rather than the production or 'audiencing' of the image (Rose, 2007, p. 61); it might not sustain comment on the cultural significance of the images made or caught. In content analysis, the whole is more than the sum of the parts, and this is particularly so in visual data, as the effect of the whole and the combination of parts can be greater than each item of composition. Content analysis must catch both the detail and the bigger picture in order to avoid being over-reductionist (Snell, 2011). As part of content analysis, coding risks losing this wholeness, being atomistic and fragmentizing. Rose (2007, p. 72) argues that content analysis does not discriminate between weaker and stronger instances of the code, and can lose important interconnections between elements of an image. Further, codes miss the mood that an image might be trying to create. Indeed she argues that, fundamentally, they overlook the important point that different people view images in different ways and with different interpretations. Whilst content analysis, conducted through coding, lends itself to the scientifically approved maxim of replicability, this may miss important features of the researcher working with visual data.

In summary, then, content analysis risks overlooking an ideology-critical way of viewing an image; it builds out such an approach, and yet ideology critique is an important element of deconstructing a visual image. Ideology, defined as the views of the ruling, dominant groups who succeed – by force or by consent (hegemony) – in having their views and values 'count' or seen as legitimate, is all-powerfully pervasive, and the views and values of others are relegated or discredited, i.e. ideology serves to reproduce social inequalities in society and to have those social inequalities played out in the everyday lives of participants. Ideology is 'lived experience', legitimating the power of the powerful at the expense of the powerless.

Ideology critique is a powerful way of looking at visual data, exposing illegitimate operations and

functions of power, and how these are produced and reproduced through images, how images legitimize social inequality (e.g. Kaufmann, 2011). This takes place, for example, in the selection, focus, exclusion, inclusion and interpretation of images and their contents. This is evidenced in semiological studies (studies of signs – signifiers – and the meaning given to that which they signify – the signified – for the viewer of the image), how meaning is encoded in the image and decoded by the viewer. In this context it is interesting for researchers to look at school prospectuses and websites; for example look at the images on the front page of school websites (e.g. Eton College and Winchester College, both of them private schools for the privileged) to see the images of the school that are selected, given or received, by the school and the viewer, to see what the images denote or connote.

Content analysis is a useful way of examining images, but its limitations have to be recognized. That is not to say that the outcomes of content analysis cannot be subject to ideology critique (indeed the study by Lutz and Collins (1993) is an example of this).

36.3 Discourse analysis

Visual images can also be read as discourses, and here the discussion of discourse analysis in Chapter 35 can apply, as images can be 'read' for the meanings that they convey to, or elicit from, the viewer. A discourse, as Rose (2007, p. 142) remarks, is a group of statements which structure how we think about things and how we act on the basis of those thoughts. As Chapter 35 makes clear, discourses structure and define what is valuable knowledge, how to know and how to think (cf. Foucault's (1998) view that discourse is an instrument and an effect of power). Discourses, like ideology, are saturated by power; hence in understanding images we have to engage in an analysis and critique of power, how it operates and with what effects (a worked example of this is presented below in an analysis of a photograph).

Discourse, as Rose (2007, p. 146) remarks, operates in several spheres, be they individual (the viewer or the producer of the image) or institutional (the items that galleries, museums etc. hold and display and how they present them). We can 'read' images for their symbolism, their messages and their iconography. This may involve trying to set on one side our own interpretations or views, and endeavouring to see the image as it might have been intended by the producer of the image, to look at the image anew, to review and review again the image iteratively and reiteratively, as Rose (2007, p. 157) remarks, to immerse ourselves in the image.

One can review the image on the basis of the structured approach of content analysis, to discover key themes or features, to identify interesting features or messages, to look for contradictions, discontinuities or complex issues in the image, to look at what the image has omitted (deliberately or not), i.e. to consider silences and absences as well as the items that have been included. In conducting this kind of discourse analysis, as with the conversational analysis in Chapter 35, there is a high level of detail in the focus and the analysis. Writing up a discourse analysis can be done through the construction of a narrative. Further, one can consider the purpose of the image in terms of its effects on the audience – intended audience or unintended audience, intended effects or unintended effects. This engages consideration of the production of the image as well as the audience of the image.

As discourse analysis and the interpretation of images involve a large element of subjectivity as intrinsic to the activity, it is incumbent on the researcher to be reflexive in the account given, indeed to regard his or her interpretation as itself a discourse.

Discourse analysis is conducted at the level of the individual image, but also at the level of the institution which holds the image, for example. the gallery, the museum, the newspaper, the film archive, the school, the broadcasting network. Rose (2007, p. 175) particularly cites this in her examples of photographs, where the home of the image may be giving messages about the institution and its values and, indeed, the intended message behind the institution's selection and use of the image, not least because institutions are sites of the operations of power (a central feature of discourses) in deciding what visual images to display or to give, together with considerations of to whom, how and where to display the images. Were the images commissioned, bought, donated, acquired, and from whom – families, philanthropists, other institutions, and how and why, and so on? How did they change hands? Here we can consider the near-instantaneous transfer of digital images in contrast to the protracted transfer of many valuable oil paintings. Images have their own social lives and biographies. What labels and captions accompany the image (e.g. the painting, the photograph), and what does it say about the priorities that the institution gives to the image? How are images stored, labelled, catalogued, archived and indexed? What are the visitor rules that must be obeyed in the viewing institution (e.g. no touching, no approaching the image too closely, no eating, no talking, no undesirable clothing (if the image is in a place of worship), how and in what order to move around the institution, where to sit etc.)?

In terms of moving images, the researcher can investigate the kinds of films that come out of film companies and studios, the kinds of programmes that television channels put out, for whom and in what format. For example, the easy-going, familiar, polite, superficial and chatty style of television talk shows, which always end on a happy note and take pains not to touch on sensitive or dangerous knowledge, can be contrasted to the gritty documentary about child prostitution or the raw film genre (Cormack, 1992). Here 'audiencing' features large: examining which audiences watch which films or which programmes, or go to see which images and where. In educational research the techniques of discourse analysis can be applied to still and moving images taken by, or provided by, the researcher and/or the participants.

Discourses and discourse analysis can apply to artefacts as well as to images. For example Francis (2010) analysed the discourses of gendered worlds into which young boys and girls are inducted through commercially produced toys and films.

36.4 Grounded theory

Both the tools and the outcomes of grounded theory can be used in analysing images. The tools of grounded theory, discussed in Chapter 37, include induction, open coding, axial coding (relating conceptually similar codes to a code that embraces them all), selective coding (looking at relationships between axial codes), categorizing, theoretical sampling, constant comparison, memoing, generation of core categories, theoretical saturation, and the generation of the theory itself as the end point of the analysis (i.e. derived from the data, not driving the data). We refer the reader to Chapter 37 for a fuller overview of these techniques. The researcher gathers together the visual data, then codes the data, moving to generating categories, themes, key issues and features, all accompanied by the writing of memos about these, thence to formulating general concepts, thence to saturating the category and theoretical sampling and onwards to the generation of the grounded theory itself. For a worked example of this with visual images, we refer the reader to Konecki (2009).

Figueroa (2008) argues that, although there is a large battery of analytical tools available for qualitative data analysis, these tend to focus on interactional studies. She argues for a variant of grounded theory to be used in analysing audio-visual texts, in the context of looking at audio-visual texts and narratives in their own right (as phenomena themselves) rather than solely regarding the audio-visual medium as the means for collecting data on a phenomenon. Texts, she avers, are 'crystallised pieces of this symbolic social net of meanings' (p. 4) and have to be examined in their own right. This entails looking at the actors' behaviours and strategies, and the consequences of these. But who are the actors – the people who have been filmed or the producers of the final image? Regarding audio-visual media simply as the means or instruments for observing a phenomenon will look at actors' behaviours and interactions; however, she suggests that it is not always easy to identify who the actors are. For example, in a piece of television journalism, the actors may be the cameraman, the journalist in the film, the chief editor, the television presenter, eyewitnesses or other people in the film, the film editor or, indeed, others. Hence it is not always easy to see who is 'speaking' in the text.

Given this difficulty, Figueroa (2008) argues that researchers have to look at texts in their own right as a single product, to see the text as a single-perspective narrative. If the researcher regards texts as the medium/means to another end, rather than as the product in itself, then this will lead the researcher to look at individual actors and their different behaviours, interactions, strategies etc. However, if texts are regarded as ends in themselves, then they will be analysed and coded differently, and, not least, interrogated for what they omit as well as what they include. Such texts and their associated readings are recognized to be: (a) already selective (having *created* a world, not only *reflected* one); (b) fictional (because they are constructed narratives); and (c) affected by the manner of their construction (they are dramaturgical and framed in a certain way, e.g. by news editors and news presenters) (p. 6).

Reading audio-visual products as texts, to be analysed through grounded theory, Figueroa suggests (p. 7), risks breaking down elements into smaller 'microscopic' units of coded fragments too soon, usually at the beginning of the analysis. This, she argues, can lose sight of the whole text and the force of the whole text, in which that whole is more than the sum of its parts. She comments that early coding analysis loses the impact of the whole when it is undertaken before any 'deep interpretation' and analysis of the overall structure of the text has been conducted.

Hence Figueroa suggests that, whilst grounded theory of texts (as products rather than as media for studying other phenomena) is useful, it should be undertaken differently from the normal sequence of open coding moving to axial coding and categorizing and, through constant comparison and the generation of core categories, to the generation of the grounded theory. Rather, she suggests that the researcher needs to

turn this approach to grounded theory on its head (p. 8). Here an analysis of audio-visual texts should start by looking at the whole, with the overall 'global impressions' and picture, as these influence the more detailed analysis that can follow. Only after the overall impression has been formed should the researcher move to the more detailed analysis and coding, i.e. with the overall picture in mind, together with an insight into the interconnections and interrelationships between different parts of the text. This echoes our earlier comment that researchers have to immerse themselves in the data and review them many times before commencing a more formal analysis, in order to see the whole picture as well as its constituent elements.

Resonant with Blumer's (1969, p. 41) advocacy of moving from the broad view to a sharper, close-up focus, this recognizes that the text is not simply a collection of independent, coded units but a whole, which has a structure and overall impact. The textual analysis becomes an 'exploration' (Figueroa, 2008, p. 9) to create a comprehensive overall picture and account of what is 'going on' in the audio-visual text, rather than simply being a coding exercise. To accompany such 'exploration', she argues for Blumer's (1969, p. 43) use of 'inspection': 'an intensive focused examination of the empirical content of whatever analytical elements' (p. 43) arise from, and come out in, the text, i.e. smaller units and pieces of the text. Indeed she writes that a more suitable way of interpreting Blumer's 'inspection' is not as examination of analytical units, but as 'exemplification' of analytical elements and emergent constructs and hypotheses.

In moving from the global to the detailed levels, macro to micro, the emergent hypotheses which are a feature of grounded theory take account of the audio-visual texts as a whole and are exemplified in the text, enabling the researcher to come to the close-up focus more slowly, after undertaking an overall view (Figueroa, 2008, p. 10). This, Figueroa avers, does greater justice to the nature of audio-visual texts and the structures of meaning within them. Though her comments are intended to apply to audio-visual texts of moving images, they can apply equally well to still images and visual data.

In advocating grounded theory, then, the researcher can start with the overall, general impression and awareness of the broad-based structures and interlocking elements of the whole, then move to the fine-grained, micro-analysis in coding and then through the several stages of the generation of the grounded theory, informed and influence by the overall impression and messages gained at the early stages of approaching the analysis.

36.5 Interpreting images

Images are 'compressed performances' (Pinney, 2004, p. 8), they take place in a social milieu, at the sites of both production and 'consumption', and the sites of 'consumption' (viewing) may change over time. They are produced for one set of purposes but often used for other purposes. The researcher has to be alert not to over-interpret photographs or to read into them meanings which are barely supportable by the material itself, i.e. he or she needs to be highly reflexive. In this respect, educational researchers should accompany the photograph in question with text, for verification, for contextualizing the photograph and, not least, for third-party validation of interpretations (ensuring that the photograph is not 'read' in entirely different ways from those of the researcher; see the comments about Figure 36.1 at the start of this chapter).

In examining images, several questions can be asked (cf. Rose, 2007, pp. 258–9):

- Why, when, where, by whom, for whom, how is/was the image made?
- Who is/was/are/were the originally intended audiences of the image?
- How is/was the image displayed?
- What do we know about the maker, the owner(s) and the people (if any) on the image?
- What were the relations (if any) between the producer, the subjects and the owner(s) of the image?
- What is the image about, and what/whom does the image show?
- What are the features of the image (e.g. compositional, genre, style, colour, elements, structure, format, arrangement, symmetry etc.)?
- What is the medium of the image?
- What are the striking features of the image?
- Is the image 'stand-alone', is it part of a set or series, is it part of a collection?
- Should the image be seen on its own or in the context of a set or series?
- From where was the image taken?
- What do the different elements of the image signify, and how do we know?
- What interpretations can be made of the image?
- Do the interpretations made of the image accord with the intentions of the producer of the image (do we know the original intentions)?
- What different interpretations of the image are made by different audiences (and from different backgrounds, e.g. related to ethnicity, age group, sex, sexuality, social class, income groups, geographical location, etc.)?

- What and whose knowledge is included in or excluded from the image?
- Who is empowered/disempowered in or by the image?
- What contradictions, if any, exist within the image?
- Where is the image kept/stored/displayed?
- Who has/had access to the image?
- How can/could the image be viewed?
- How is the image described, labelled, indexed, catalogued, archived?
- Is there a written commentary on the image, and, if so, what does it contain?
- What is the intended and actual relation between the image and those who view it?

There is a wealth of literature on examining images in educational research, particularly in the history of education, and we refer readers to O'Donoghue (2010) for comprehensive references here. His paper also suggests that images, including photographs, can be regarded as 'installation art', i.e. those artworks that are produced at an exhibition site. Regarding photographs as 'photographs of installations' (p. 411) invites researchers to imagine not only the three-dimensional nature of the classroom but also how it must feel to be inside that classroom.

36.6 Interpreting an image: a worked example

A worked example of a 'reading' of an image is presented Figure 36.2. This is a still image, a photograph.

This fascinating historical photograph of a UK schoolroom in the north-east of England carries the museum label thus: 'Children possibly at Woodland school, taken during an art class. Note sculptured trees on desks.' It is a typical photograph of its time (early twentieth century), and, indeed it is part of the genre of this type of photograph in which each child's head is turned to the left, the teacher is at the back of the class and the photographer is on one side of the room in order to include all the children in the photograph (and

FIGURE 36.2 An early twentieth century photograph of children in an art lesson
Source: Image courtesy of Beamish Museum, image copyright Beamish Museum

maybe to utilize natural light). In places the photograph is faded and the image is a little fuzzy: the ravages of time and technology. It has also been preserved in digital form by the museum, so that further image quality loss is prevented.

If we examine the picture, what can we notice?

The people

- There are sixty children in the class (there may have been just a few more, out of the camera shot on the right; the presence of light from the right suggests that the last row on the right may be next to a window).
- There are more girls than boys.
- The sexes sit together, and indeed in some places a boy is wedged between two girls.
- All the children are white Caucasians.
- The teacher is female.
- Nearly all of the children are dressed smartly in the style of the day; it is unclear whether there is a uniform, or clothing for the special event of the photograph, but there is a homogeneity or standardization of clothing.
- Some boys are wearing expensive lace collars, others are wearing stiff 'Eton' collars, but the school is probably not for rich children (who would be in much smaller classes and with different uniforms; perhaps here the parents wanted the best for their children).
- Clothing is clearly differentiated by sex.
- The children are wearing warm clothing.
- The only person not looking at the camera is the teacher, and, like a military officer, she is looking imperiously, sternly and unsmilingly at the children, and is the only one standing in the photograph, i.e. physically and metaphorically above the students.
- All the children are facing the camera; no child is looking away.
- The picture is 'posed' and serious, not light-hearted; clearly the children have been told what to do, how to sit (hands behind their backs) and where to look. Some are trying to smile, one or two seem to be smiling more naturally, and yet most are not.
- The situation seems unusual for the children, to have a photographer in the classroom, as many of them have an air of curiosity in their look.

The classroom and the furniture

- Proportional to the number of people, the classroom is quite small and the children are tightly packed.
- The back of the classroom is raised up (by one step, visible on the upper right of the photograph), so that the children at the back can see the teacher at the front, and be seen by that teacher.
- There are no windows out of which children can look (the windows are too high or are blocked out).
- The children are sitting in solid desks, three to a desk.
- The desks are standardized, the same, dark (black iron and dark wood), heavy (too heavy to move easily) and unable to be adjusted.
- The desks are large, taking up all the classroom space, yet the children are small. The desks are bigger than the children.
- The desks are fixed, made of strong wood and cast iron.
- The desks are hard, strong and large, in contrast to the students who are fragile and small.
- There is little room for movement in the desks; the position of the seats is fixed, as they are joined to the desk by the iron bar at the base.
- The seating arrangement suggests that all the interactions go through the teacher.
- The seating arrangements may be designed to control children, not least the boys (mixing the sexes and having some boys sitting between two girls).
- The children sit in rows and columns, each row facing the front. It is very regimented, and oriented to a single focal point – the teacher at the front.
- There appears to be a gap, a distance, between the front row of children and the teacher's desk (out of the image).
- There are some unusual objects in the class: the large thermometer hanging from the light fitting (a science instrument?), the large portraits high up around the room (not all completely contained within the photograph), with dignitaries looking down on the children.
- There is bare, but varnished, brickwork in the classroom.
- Some work that is on the walls is too high for children to read – it is for decoration only.
- The children's pictures are nearly all the same, and are about the same topic – flowers; all are nearly identical.
- The pictures by the children, on the walls, are stylized and almost the same.
- There is an almost exclusive focus on nature in the children's pictures and no other work is on display (indicative, perhaps, of an alternative to the hardness of the real world inside and outside the classroom).
- This is an art lesson, yet there is no evidence of drawing materials. There is evidence of what the children should be looking at in the art lesson (the

jar of flowers on their desk or the sculptured trees). It is unclear whether this is an art/drawing lesson or an art appreciation lesson.
- All the objects on the desks are the same.

The photograph and the photographer

- The photograph is old, and, in parts, the focus is not always sharp or even, the images are slightly unclear in places, the contrast is uneven and, in parts, the image is faded. Hence the researcher has to be careful not to over-interpret those parts of the photograph which are unclear or to read into the analysis any points that are not supportable by the evidence. This is a commonplace problem with old materials, and argues for the value of a third party to examine the photograph.
- The photographer must have been standing some distance from the children (nearly two desks' length from the front row of desks if we calculate the ratios) and higher than floor level. Standing higher than the children makes them look smaller – the symbolism is striking.
- The way in which, taken as a two-dimensional image, the teacher is at the apex and the children are below, constitutes a visual hierarchy reflecting a positional/role hierarchy.
- Why was the picture taken? For whom? For what purpose?
- There is no clear single focal point in the photograph; the conventional 'rule of the thirds' (where the focus is one third or two thirds of the way into the picture) is not there, nor is there a clear centre to the image.
- There are many points of focus, for example:

  a the girls' bright dresses in the first complete right-hand row;
  b the staring eyes of the boy sitting at the front, or the worried look of the little girl in the second row, or the haughty teacher at the back;
  c the children who are more in the image's sharp focus towards the rear of the second row of desks;
  d the bright lace collar of the boy in the centre rear;
  e the near-rhomboid symmetry in terms of the rows and columns of children's heads, which suggests order, regulation and regularity;
  f the use of diagonals here, rather than a front shot (whether simply out of the requirements to include all the children seated in their desks, or for artistic effect, or to make the most of the natural light, or some other reason), which brings a sense of inclusiveness to the picture and which draws the viewer into the picture;
  g the field of vision of the viewer (from a single point outwards), which is matched by the shape of the classroom (Figure 36.3) and the view of the arrangement of the desks and children (almost a rhombus, see Figure 36.3);
  h the match between the direction of the walls of the classroom and the layout of the rows and columns of the desks (the children are triply 'contained': (a) within their desk; (b) within the rows and columns of the desk arrangement; and (c) within the confines of the classroom walls, all of which is supervised by the overriding presence of the teacher). There is a scalability to the picture: each desk is a scaled-down version of the arrangement of all the desks (into rows and columns) and the arrangement of all the desks is a scaled down version of the proportions and layout of the classroom walls;
  i the contrast between the foreground and the background – the foreground shows powerless children whilst the background shows the powerful teacher keeping watch;
  j the dowdy walls and gloomy far reaches of the classroom contrast with the humanity and clothed children and models sitting in the centre of the picture;
  k the contrast between the harsh brick walls and the soft children;
  l the contrast between the staid and very formally dressed teacher and the relatively innocent children's faces and clothing;
  m the emphasis on regularity (rows and columns) and the repeated motifs of the three children sitting at a desk, multiplied eighteen times (eighteen complete desks in the picture);
  n the contrast between the static pose rather than the dynamic potential of the photograph, there being sixty-one potentially dynamic agents (people) in the photograph.

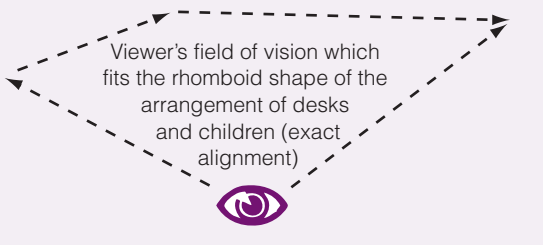


FIGURE 36.3 Matching the viewer's field of vision and the shape of the main part of a photograph

- The way in which the picture's background is cut off at crucial points.
- The old, faded and fuzzy parts of photograph.
- The observation that there are almost no shadows, everything is open to scrutiny and nothing is shaded or hidden.

What we see is often what we look for; this makes us look selectively and construe what we see through the interpretive lenses of our own subjectivity and ideological frameworks and values. Researchers bring their own subjectivities and cultural backgrounds to the photograph (hence the issues of reflexivity and disclosure of possible subjectivity assume a high profile here).

For example, one researcher might 'read' this picture as presenting stark messages and themes:

- lack of freedom and no room for freedom;
- power (the teacher has it all and the children seem to have none): asymmetrical relations of power;
- lack of creativity;
- standardization, sameness and uniformity;
- surveillance, control, domination, authoritarianism and containment;
- the gendered nature of primary school teaching;
- conformity, obedience, passivity and loss of individuality.

Though the formal curriculum here may be art (which, perhaps, concerns individuality and creativity), the hidden curriculum (that which is learnt without being taught; the unspoken messages that children must learn very thoroughly if they are to survive in school, e.g. about being one of a crowd, about differentials of power, about delay, denial and domination) (Jackson, 1968) is the exact opposite.

Of course, this interpretation might say more about the researcher than the researched: the researcher may be attuned to looking for dominatory forms of schooling, to the neglect of its more positive aspects. For example, another researcher may interpret the photograph as showing:

- a clear, undistracted focus on the teacher and children's own work, designed to promote learning and concentration;
- clear understanding by all parties of roles and behaviours, so that learning can take place beneficially, willingly and without disruption.

Here the researcher may feel that the clarity of role specifications and expected behaviours are not at all negative, but are designed to promote the effective learning of children. Indeed the power of this arrangement for learning and its outcomes could be immense, for example, for children to be able to climb the social ladder in the future: education as a great emancipatory force in society.

Further, initially we have the photograph's title attached to it by the museum: 'School Children in Art Class', with the museum's own label reading: 'Children possibly at Woodland school taken during an art class. Note sculptured trees on desks'. Immediately the reader's attention is drawn to the fact that this concerns an art lesson, and that there are some art materials. Why were these features included in the text, and not others? Is that really the purpose or key message of the image, or is the museum, in a positive endeavour to be helpful, drawing attention to points that otherwise might go unnoticed? Is it trying not to be pejorative in its comments, or is it simply that the museum wanted a short label for indexing and referencing purposes? The point here is that labels can frame the researcher's or the viewer's insights, and the researcher needs to be aware of this. It is not only the focus of the text label, but the tone of those words: the label used by the museum may appear to be couched in neutral terms, but it has already decided what to comment on and what to ignore. Guidelines on inclusion and exclusion can be both useful and dangerous.

In considering the photograph, indeed any visual image, we can focus on the subject matter, its form, its genre, its meanings, its composition, its style and technical matters. However, we can go further, to examine the context of the photograph, its audience, its provenance, why it was taken, its usages and, indeed, the ethical issues that are raised by the photograph.

The researcher can speculate on the history of the photograph in question: why it was taken, for whom it was taken and what use was intended to be made of it, or indeed was made of it? Was it designed to impress parents, school governors, inspectors, local officials (and, if so, why was an art lesson chosen)? Was it designed to be simply a document of record of the school's history, and if so, why this scene in particular? Was it designed to be a celebratory record (the children may have been dressed smartly for the occasion, in clothes that they would not normally wear for school)? Who was the intended audience: the children themselves (e.g. in later life), their parents, education officials, researchers, visitors, historians, the families in question?

We can also ask how and why the photograph came to be in the museum in question (an award-winning national museum of social and industrial history).

For example, was it a donation, a purchase, did it arrive by happenstance, or deliberately, or as part of a large collection, or what?

The photograph raises several ethical questions, for example:

- Are the people still alive?
- Was informed consent gained from the people in the photograph to be photographed (or was it simply an accepted part of being at school)?
- What informed consent was gained by the museum, and from whom, to release the document into the public domain, or has the passage of time obviated the need for this?
- Is it acceptable and fair of the researcher to portray the school, the teacher and the students in question in a perhaps negative way, and, if not, then who actually suffers?
- Will the use of the photograph bring harm or good, and to whom?

What we have here encapsulates the problem that often adheres to documentary evidence: that it is prepared (or in this case taken) for one set of purposes and audiences, but it is used for other reasons and intentions.

Photographs, like other visual materials, are multilayered and capable of sustaining several interpretations. Hence the visual researcher, just like the textual researcher, has to disclose his or her own reflexivity and the possible influence that this has on the analysis and interpretation made. Though a picture may be worth a thousand words, photographs on their own may be relatively inert; it is only in the interaction between the producer of the image, the image itself and the audience that it comes alive.

We can read a photograph like a text, and indeed it is often useful to accompany the image with text. Text and photograph run together. A commentary can be useful to accompany, explain, interpret and contextualize the image, and, in research terms, this can tie the image into other evidence – visual or textual – that the researcher is using.

36.7 Analysing moving images

The term 'moving images' here is taken to include video and film material. Denzin (1990, p. 102) remarks that 'films do not faithfully reproduce reality'; rather, they are ideological interpretations and selections from reality; they are a particular version or view of reality. Hence the researcher has to interrogate the moving images in light of the research questions and to undertake a more valuative and ideology-critical reading of their content. This includes selecting, and justifying the selection of, particular parts of the moving images (what to focus on and what to overlook), which may be informed by the research questions and purposes.

Denzin (2004) suggests that films (including videos) should be considered initially at their 'textual realism' level, i.e. the story that the material is telling and how it is telling that story. At a second level, which he terms a 'subversive' level (p. 240), he suggests that a film can be read for its ideological content and effects, i.e. how the film functions to reproduce the (dominant) values and beliefs of everyday life and society. Hence the researcher starts with an overall view of the film as a whole, noting themes, impressions, key points, rather as one would 'read' a text. Having gained an overall view, the researcher can then go into details, for example, scenes, events, sequences and so on, in short, a micro-analysis of the material (Flick, 2009, p. 247). In this, the methods and tools of grounded theory, discourse analysis etc. can be used, working not only with the visual images but also, where relevant, transcriptions of the spoken words. As Flick remarks (p. 249), films can be regarded as visual texts, and so the range of tools for textual analysis can be brought into play here. He argues that researchers can look for patterns in the film (p. 247). Having conducted a first-level and second-level analysis, the researcher can then look for points of resonance, consonance, dissonance and contradiction between the two levels of analysis.

In video analysis, researchers can decide the unit of analysis (e.g. a time interval, a setting, an event, a sequence, an interaction) (Lee *et al*., 2015) and then analyse data using those units of analysis. The researcher can use inductive approaches (without a hypothesis or theory to be tested) and deductive approaches (hypothesis testing), time sampling or event sampling of video footage, both random and purposive, and looking for similarities and differences between coded elements (Lee *et al*. 2015). As in other forms of data analysis, coding can be used (Johnson and Black, 2012), for example, for quantitizing qualitative data or simply as part of conventional coding and analytical practices, indeed Lee *et al*. (2015) report using colour coding to indicate levels of interaction in classrooms, from cold colours for limited interaction to warm colours for greater interaction.

In selecting the time interval for coding the researcher has to balance too fine a level of detail, which might overlook the big picture, with too broad a picture, which might overlook important detail (Snell, 2011; Lee *et al*., 2015). Snell (2011), using the software The Observer XT, comments on its possibilities for generating systematic observation of video data

for statistical analysis, and for micro-ethnographic analysis, using coding and data management of video data. Codes and their related data are time-stamped, logged and saved in dedicated locations, synchronizing the codes and their related video data. Snell makes the point that detailed analysis contributes to nuanced interpretation and prevents over-reduction to which systematic analysis might be prone. Conversely, systematic analysis might overcome the 'cherry picking' of video clips and the overlooking of the bigger picture (p. 257).

As with much qualitative data analysis, exploration and interpretation run together; hence researchers have to be acutely aware of the influence of their own values, cultures, interests and background in the selection and interpretation of the data; in short they have to be reflexive. In this respect the repeatability of moving image material is useful in being able to be viewed by a third party, to check for alternative interpretations of the material.

Analysing moving images is costly in terms of time, as they have to be watched and re-watched many times in order to extract fair and suitable data and interpretations (e.g. for coding, constant comparison and narrative construction). This can be exacting and demanding of the researcher's insight and persistence.

36.8 Conclusion

This chapter has suggested that analysis and interpretation of images are often inextricably linked, raising the need for considerable reflexivity on the part of the researcher. Content analysis (both numerical and qualitative), discourse analysis and grounded theory can be used in the analysis and interpretation of visual images. The chapter provided a worked example of the analysis and interpretation of a single still image: a photograph. It has used this not only to indicate the processes and kinds of observations and interpretations that can be made, but to indicate how interpretations are multiple, sometimes conflicting, and subjective. The interpretation used elements of ideology critique in its exposure and disclosure of power in the image and its explanation. The authority of the researcher to determine the focus, analysis and interpretation of a still or moving image is, itself, subject to ideology critique and interrogation of power within a discourse. Visual images invite researchers to consider alternative forms of representing data, alternative questions to be asked about education systems and new ways of looking at and seeing things (Eisner, 1997, p. 6). That is a powerful challenge.

Companion Website

The companion website to the book provides PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Grounded theory

CHAPTER 37

One intention or outcome of analysing qualitative data is the generation of theory. Here we introduce key issues in grounded theory and the tools of grounded theory. The chapter includes:

- versions of grounded theory
- stages in generating a grounded theory
- the tools of grounded theory
- the strength of the grounded theory
- evaluating grounded theory
- preparing to work in grounded theory
- some concerns about grounded theory

Readers may find it helpful to refer also to Chapter 15 and particularly Chapter 34 on coding, as coding is integral to grounded theory (see below: coding).

37.1 Introduction

Theory generation in qualitative data can be emergent, and grounded theory is an important method of systematic emergent theory generation. Strauss and Corbin (1994, p. 273) remark that grounded theory is a methodology which seeks to develop theory which is rooted in – grounded in – data which have been collected systematically and analysed systematically; it is an orderly, methodical and partially controlled way of moving from data to theory, with 'clearly specified analytical procedures' (Greckhamer and Koro-Ljungberg, 2005, p. 731). It is more inductive than content analysis as, typically though not exclusively, the theory emerges from, rather than exists before, the data. The theory is derived inductively and, in some versions, deductively and abductively, from the analysis and study of, and reflection on, the phenomena under scrutiny (cf. Strauss and Corbin, 1990, p. 23).

Grounded theory, as Moghaddam (2006) avers, is a set of relationships among data and categories that proposes a plausible and reasonable explanation of the phenomenon under study, i.e. it explains by drawing on the data generated. It is a method or set of procedures for the generation of theory or for the production of a certain kind of knowledge (Greckhamer and Koro-Ljungberg, 2005, p. 729; Kolb, 2012). Glaser, one of the key writers in grounded theory, focuses on the methods of grounded theory, and Birks and Mills (2015, p. 5) add to this its philosophical roots in pragmatism and symbolic interactionism.

Grounded theory uses systematized methods (discussed below) of theoretical sampling, coding and categorization, constant comparison, memoing, the identification of a core variable, and saturation, all of which lead to theory generation. Grounded theory is not averse to quantitative methods, indeed it arose out of them (Glaser, 1996) in terms of trying to bring to qualitative data some of the analytic methods applied in statistical techniques (e.g. multivariate analysis). In grounded theory the researcher discovers what is relevant; indeed Glaser's and Strauss's (1967) seminal work is entitled *The Discovery of Grounded Theory*.

However, where it parts company with much quantitative, positivist research is in its view of theory. In positivist research the theory pre-exists its testing and the researcher deduces from the data whether the theory is robust and can be confirmed. Grounded theory, on the other hand, does not force data to fit a predetermined theory (Glaser and Strauss, 1967, p. 3); indeed it eschews such 'forcing', though the difference between inductive and deductive research is less clear than appears at first sight. For example, before one can deduce, one has to generate theory and categories inductively. The intention of grounded theory is to build and generate theory rather than to test an existing theory, providing researchers with tools to generate this theory through data analysis, to weigh up alternative explanations (e.g. through constant comparison) and to relate concepts in the development of theory (Moghaddam, 2006; Birks and Mills, 2015).

Grounded theory starts with data which are then analysed and reviewed to enable the theory to be generated from them; it is rooted in the data and little else. Theory derives from the data; it is grounded in the data and emerges from it. Glaser (1996) writes that 'forcing methodologies were too ascendant', not least in positivist research, and that grounded theory had to reject forcing or constraining the nature of a research

investigation by pre-existing theories. As grounded theory sets aside any preconceived ideas, letting the data themselves give rise to the theory, certain abilities are required of the researcher, for example:

- tolerance and openness to data and what is emerging;
- tolerance of confusion and regression (feeling stupid when the theory does not become immediately obvious);
- resistance to premature formulation of theory;
- ability to pay close attention to data;
- willingness to engage in the process of theory generation rather than theory testing; it is an experiential methodology;
- ability to work with emergent categories rather than preconceived or received categories.

As theory is not predetermined, the role of targeted prereading of literature is not as strong as in other kinds of research which use literature reviews to generate issues for the research. Indeed conducting prior literature reviews may be dangerous as it may prematurely close off or determine what one sees in data; it may cause one to read data through given lenses rather than with fresh eyes. As one does not know what one will find, one cannot be sure what one should read before undertaking grounded theory. One should read widely within and outside the field, or not at all, rather than narrowly and in too focused a direction (Glaser, 1996).

37.2 Versions of grounded theory

There are many versions of grounded theory (Hutchison *et al*., 2010), and indeed Greckhamer and Koro-Ljungberg (2005) comment that what grounded theory actually is has become a contested issue (p. 731). Three widely referenced versions are:

- the original, emergent model by Glaser and Strauss (1967) and Glaser (1978, 1992);
- the revised, systematic model by Strauss and Corbin (1990, 1998) and Corbin and Strauss (2008, 2015);
- the constructivist model by Charmaz (2006).

They differ in their processes, key elements, epistemologies, ontologies, theoretical foundations and frameworks. There are other models (e.g. Clarke's (2007) situational analysis model) which we do not include here.

The original, emergent model

In the original model articulated by Glaser and Strauss (1967), the theory emerges from the data, using various tools to facilitate such emergence and the 'discovery' of the theory that is embedded in the data. In this process there are two main types of coding: substantive and theoretical; substantive coding (with open coding) precedes theoretical coding (with selective coding and theoretical sampling).

This model uses memoing, constant comparison, theoretical sampling, derivation of the core category and theoretical sampling (all discussed below). Here the authors argue against starting with a literature review as this could lead to preconceived ideas which could influence too strongly or bias the researcher's responsiveness to the data, as, in their view, the researcher must be as open to the data as possible (Glaser and Strauss, 1967; Glaser, 1998). Substantive codes deliberately 'fracture' the data whilst theoretical codes deliberately recombine them into an organized whole.

The revised, systematic model

The revised grounded theory model from Strauss and Corbin (1990, 1998) is much more systematic and prescriptive than the original model, so much so that there was a well-documented split between Strauss and Glaser, with Glaser arguing that the revised model of Strauss and Corbin was formulaic and too prescriptive, 'forcing' a theory onto data and forcing data into a theory, whereas the essence of grounded theory was its aversion to forcing. (The third and fourth editions of their book (Corbin and Strauss, 2008, 2015) are more flexible and accommodating than the first edition in terms of processes and procedures.) The revised model gave prominence to axial coding (axial coding was absent from the original model of Glaser and Strauss (1967)), with a prescribed sequence of open coding leading to 'axial coding' and 'axial coding' leading to 'selective coding'. The analytical process was more prescriptive, detailed and predetermined than the original, emergent model, and this new model included a required 'conditional matrix' which was not present in the original model.

This new model also accorded greater value to the literature review early on in the research process (on the grounds that there is no *tabula rasa* in the mind of the researcher), as this can develop theoretical sensitivity and hypothesis generation, whereas the original model deliberately argued against researchers conducting a literature review in advance. Further, the revised model adopts a more prescriptive approach to the types of memos researchers write (e.g. operational, coding-related, theoretical), whereas the original model kept the nature of the memo open and flexible. The original model emphasized an inductive approach to data analysis whereas the revised model by Strauss and Corbin

included deductive approaches and theory verification by the data.

The Strauss and Corbin model differs from the original model as they argue that: (a) sampling proceeds on theoretical grounds (a point which the original model rejects as introducing bias into the research and that the research should commence with data alone); (b) hypotheses can be developed and verified (whereas the original model abjures prior theory generation, testing and verification); (c) induction, deduction and abduction can be used, whereas the original model advocates induction alone and rejects deduction; and (d) attention must be given to broader structural contents and influences (whereas the original model argues that these, if they are present in the research, would be manifested in the data alone and, otherwise, should not be considered).

The constructivist model

In the constructivist model from Charmaz (2006), subjective meanings are attributed to the data by participants and researchers, and there might be multiple interpretations of what these meanings are. This moves beyond 'facts' and descriptions of acts to interpretations and perspectives. Charmaz holds that concepts are not so much revealed or 'discovered' (the title of Glaser's and Strauss's original book (1967)) as 'constructed', for example through interactions and involvements, both past and present, with people, ways of looking, interpretations and meanings, leading to one or more 'constructions of reality' (p. 10).

The theoretical basis of her view, rather than being objectivist as in the two previous models, is interactionist and constructivist (Charmaz, 2002, p. 678). Data are 'reconstructions of experience; they are not the experience itself' (Charmaz, 2000, p. 514) and hence are open to interpretation, and there may be more than one grounded theory that emerges from the data (Greckhamer and Koro-Ljungberg, 2005, p. 744).

Charmaz (2002) sets out six analytical steps in her grounded theory: (i) data collection and analysis, simultaneous and ongoing; (ii) early data analysis to identify emergent themes; (iii) identification of basic social processes from and within the data; (iv) inductive construction and co-construction of abstract explanatory categories for those processes; (v) constant comparison to refine the categories; and (vi) integrating the categories into a theoretical framework that identifies causes, consequences and conditions (cf. Greckhamer and Koro-Ljungberg, 2005, p. 739). In this constructivist model there are three types of coding: open, focused and theoretical. Theoretical coding here means the fusing together of concepts into groups as the analysis proceeds, not as a later stage or end-point of the analysis.

In the constructivist model an initial literature review is entirely acceptable. This model is less objectivist than the other two models, and emphasizes subjective constructions and co-constructions of knowledge, the generation of different meanings by participants and researchers, and the openness to modification of any emergent theory in light of those meanings and interpretations. Objective notions of reality, as discovered by a neutral observer, are replaced by the (social) construction of subjective meanings of reality (Keane, 2015).

Hernandez and Andrews (2012) contend that, whereas the original model of grounded theory gives rise to explanatory theory, the constructivist model gives rise to a descriptive theory. Indeed Charmaz (2002) notes that grounded theorists working in the constructivist perspective admit that it is they who define and *construct* what is taking place in the data, in contrast to objectivist grounded theorists who *discover* what is taking place in the data (p. 684).

The models from Strauss and Corbin (1990, 1998) and Charmaz both reject the idea of holding back from conducting a preliminary literature review. Similarly, Dunne (2011) sees it as simply unworkable for researchers seeking permission to conduct research or to obtain research funding not to demonstrate that they have done their homework and that they are familiar with the field, both of which require a literature review. There is no good reason why a researcher cannot review relevant literature, as this is part of the 'historicity', familiarization and contextualization of the study, just as other familiarization activities might be acceptable, indeed unavoidable. As Thornberg (2012a) argues, researching the literature can support greater sensitivity and creativity, spotting matters that might otherwise be overlooked.

Common features of the three models

Though there are different versions of grounded theory, and variations in its forms and epistemologies (Greckhamer and Koro-Ljungberg, 2005, p. 731; Buckley and Waring, 2009, p. 318; Hutchison *et al*., 2010; Waring, 2012), nevertheless there are several features in common in these definitions:

- theory is *emergent* rather than predefined and tested; it emerges from the data rather than *vice versa*;
- the grounded theory process is recursive;
- sampling is targeted at generating theory;
- the process of analysis involves coding and categorization;

- data collection and analysis proceed simultaneously and are ongoing;
- systematic comparisons are made in the process of analysis ('constant comparison');
- theory generation is a consequence of, and partner to, *systematic* data collection and analysis;
- patterns and theories are implicit in data, waiting to be discovered;
- grounded theory is close to the data that give rise to it.

Glaser (1996) states that 'grounded theory is the systematic generation of a theory from data'; it is an inductive process in which everything is integrated and in which data pattern themselves rather than having the researcher pattern them, as actions are integrated and interrelated with other actions. Glaser's and Strauss's (1967) seminal work rejects simple linear causality and the decontextualization of data, arguing that the world which participants inhabit is multivalent, multivariate and connected. As Glaser (1996) says: 'the world doesn't occur in a vacuum', and the researcher must take account of the interconnectedness of actions. In everyday life, actions are interconnected and people make connections naturally; it is part of everyday living, and hence grounded theory catches the naturalistic element of research and formulates it into a systematic methodology. In seeking to catch the complexity and interconnectedness of everyday actions, grounded theory is faithful to how people act; it takes account of apparent inconsistencies, contradictions, discontinuities and relatedness in actions. As Glaser says: 'grounded theory is appealing because it tends to get at exactly what's going on'. Flick (1998, p. 41) writes that the aim is to recognize complexity by including contextual details, rather than to reduced complexity by atomizing it into variables.

There are several elements of grounded theory that contribute to its systematic nature and it is to these that we now turn.

37.3 Stages in generating a grounded theory

The stages in generating a grounded theory depend, in part, on the model of grounded theory adopted, and they may vary. However, a typical sequence is:

1. Decision on whether a grounded theory approach is most suitable →
2. Theoretical sampling + memoing →
3. Data collection + memoing →
4. Coding: open codes leading to axial codes (clustering open codes into groups by meaning), leading to selective codes (relating axial codes to each other and drawing linkages); theoretical codes + memoing →
5. Categorization (which might involve reducing the number of codes and creating hierarchies of codes) + memoing →
6. Constant comparison + memoing →
7. Identification of the core variable + memoing →
8. Saturation + memoing →
9. Theory generation/verification →
10. Writing the report.

However, this suggests linearity, a sequence. Though Strauss and Corbin (1990, 1998) advocate sequencing (including adding a conditional matrix to the sequence above), in reality the researcher will operate recursively, going back and forth in these stages, or indeed operating several stages in parallel, reviewing, revising and reworking as considered appropriate.

37.4 The tools of grounded theory

There are several common tools that researchers use in grounded theory: theoretical sampling; coding (discussed in Chapter 34); memoing; constant comparison; identification of the core variable(s); 'saturation'; and theoretical sensitivity. We discuss these below.

Theoretical sampling

Theoretical sampling, as Glaser and Strauss (1967) write, is a process for generating theory. In this the researcher collects the data, processes data with coding and analyses the results, and this analysis informs where to go next in collecting data in order to develop the emerging theory (p. 45), i.e. the emerging theory controls the process of data collection and is the criterion – theoretical relevance – for proceeding further with data collection (p. 49) rather than, for example, conventional sampling approaches. Data are collected which are useful to the generation of theory (Creswell, 2012, p. 433), i.e. purposive sampling takes place.

Theoretical sampling is that kind of sampling which is based on the concepts which have shown themselves to be theoretically relevant to the evolving or emerging theory (Strauss and Corbin, 1990, p. 176; Birks and Mills, 2015, pp. 68–71). Here data are collected and analysed on an ongoing basis, with the analysis informing the further collection of data, i.e. a process of continual refining of the categories, ideas, concepts and emergent theory. The researcher keeps on adding cases to the sample until she has enough data to describe what is going on in the context or situation under study and until 'theoretical saturation' is reached (discussed

below) (Glaser and Strauss, 1967; Strauss and Corbin, 2008). As one cannot know in advance when this point will be reached, one cannot determine the sample size, nature or representativeness until one is actually doing the research. Kolb (2012) suggests that, therefore, theoretical sampling might lead to biased sampling.

In theoretical sampling, data collection continues until sufficient data have been gathered to create a theoretical explanation of what is happening and what constitutes its key features. It is not a question of representativeness, but, rather, a question of enabling the theory to emerge. Corbin and Strauss (2008) write that purposeful sampling takes place at the levels of open coding and selective coding, whilst structured, systematic sampling takes place at the level of axial coding.

Coding

Coding is the process of disassembling/fracturing and then reassembling the data. Data are disassembled when they are broken apart into lines, paragraphs or sections, subsequent to which these fragments are rearranged, usually through coding, to produce an organized and structured thematization and theory (cf. Ezzy, 2002, p. 94).

In grounded theory there are three main types of coding: *open*, *axial* and *selective*, the intention of which is to disassemble the data into manageable chunks in order to facilitate an understanding of the phenomenon in question. Strauss and Corbin (1990) suggest a linear sequence here: open coding to axial coding to selective coding. Whether such linearity is acceptable is contestable, as recursion and iteration can occur.

Open coding involves exploring the data and identifying units of analysis to code for meanings, feelings, actions, events and so on. The researcher codes the data, creating new codes and categories and sub-categories where necessary, and integrating codes where relevant until the coding is complete. *Axial coding* seeks to make links between categories and codes, 'to integrate codes around the axes of central categories' (Ezzy, 2002, p. 91); the essence of axial coding is the interconnectedness of categories (Creswell, 2012). Hence codes are explored, their interrelationships are examined and codes and categories are compared to existing theory. In *selective coding* a core code or category is identified, the relationship between that core code/category and other codes/categories is made clear (Ezzy, 2002, p. 93), and the coding scheme is compared with pre-existing theory. Creswell (1998, p. 57) writes that here the researcher identifies a 'story line' and proceeds to construct the story that draws together all the axial codes.

As coding proceeds the researcher develops concepts and makes connections between them. Flick *et al*. (2004, p. 19) argue that 'repeated coding of data leads to denser concept-based relationships and hence to a theory', i.e. that the richness of the data is included in the theoretical formulation.

We strongly advise readers here to consult Chapter 34 on coding, as it provides much more detail.

Memoing

Chapter 33 introduced memos, and here we take the matter further. Memos are simply notes written to oneself, logging ideas, abstract thoughts, insights, observations, conjectures and possibilities etc. (cf. Waring, 2012, p. 302; Denscombe, 2014, p. 285; Birks and Mills, 2015). They are typically written by the researcher to herself/himself and contain analysis which contributes to the formulation of the emergent theory (Strauss and Corbin, 1990, p. 197). They can contain notes on a wide field of matters; they can be short, long, detailed, general, focused, wide-ranging etc. They can be, for example, conceptual, theoretical, operational, reflexive and coding-related; they are what the researcher wants them to be.

In memoing, the researcher writes ideas, notes, comments, notes on surprising matters, themes or metaphors, reminders, hunches, draft hypotheses, references to literature, diagrams questions, draft theories, methodological points, personal points, suggestions for further enquiry etc. that occur to him/her during the process of constant comparison and data analysis (Lempert, 2007, p. 245; Flick, 2009, p. 434).

Hutchison *et al*. (2010) note that memos can be: *a research diary* (e.g. containing conceptual developments and general events); *reflective and reflexive; conceptual* (often attached to codes, nodes and categories); *emergent questions* and summaries of *emergent themes* and *explanatory* (e.g. related to literature, technical matters, models). Waring (2012) suggests three main types of memo: *code notes* (e.g. containing the names of codes and how these were derived); *theoretical notes* (extensions of code notes, e.g. containing the products of inductive and deductive thinking in relation to properties of, and relationships between, data, codes and theorizing); and *operational notes* (e.g. concerning the conduct of the data collection, research and data analysis) (p. 302).

Memos cover many aspects of the research and data analysis. They can address many matters, for example, those set out here in alphabetical order:

- analytical notes and ideas;
- codes, categories and the products of coding;

- comments on sampling;
- comments on saturation;
- concepts and key concepts;
- conditions and contingencies;
- conjectures and speculations;
- core category;
- cross-references and relationships;
- decisions taken;
- descriptive details;
- diagrams;
- directions and suggestions;
- emerging theory;
- explanatory ideas;
- feelings about the research;
- grounded theory;
- ideas;
- impressions;
- inductive and deductive material;
- issues and ideas arising in the research;
- models;
- observations;
- operational matters;
- philosophical matters;
- procedural matters;
- reflections on the research;
- relationships and comparisons;
- reminders;
- suggestions for further directions of investigation;
- summaries;
- themes;
- theoretical sampling;
- theoretical sensitivity;
- theoretical suggestions;
- theory.

Memos can be written at any stage of the data collection and analysis; they can vary in length and format, from informal to more formal. They may contain verbatim quotations, notes, jottings, key points underlined, diagrams (Strauss and Corbin, 1990, pp. 202–3); in short, nothing is ruled out.

Memos should bear a date, references to the data/data file about which they are written, a heading to identify what they are about and, where appropriate, references to relevant codes and categories (Strauss and Corbin, 1990, pp. 200–4). Memos, in turn, become data, and indeed memos about memos can become data (e.g. to enable reflexivity). Much software (e.g. NVivo) provides a repository and location for memos and links them to primary data sources.

Constant comparison

Constant comparison is the process 'by which the properties and categories across the data are compared continuously until no more variation occurs' (Glaser, 1996), i.e. saturation is reached. In constant comparison, data are compared across a range of situations, times and groups of people, and through a range of methods. The process resonates with the methodological notion of triangulation.

The application of open, axial and selective coding adopts the method of constant comparison. In constant comparison, the researcher compares the new data with existing categories, so that categories achieve a perfect fit with the data. If there is a poor fit between data and categories, or indeed between theory and data, then the categories and theories have to be modified until all the data are accounted for. New and emergent categories are developed in order to incorporate and accommodate data in a good fit, with no discrepant cases. Data collection and analysis proceed together.

Glaser and Strauss (1967, p. 102) note that constant comparison using coding and data analysis is the means for generating theory. The theory generated does not explicitly seek generalizability; rather the researcher seeks theoretical saturation, i.e. no further data modify the theory. Constant comparison is the process by which the properties and categories across all the data are compared continuously until no more variation is found (Glaser, 1996).

To accompany constant comparison, and to aid reflexivity, Glaser and Strauss (1967) suggest the value of memoing, discussed earlier. In constant comparison, discrepant, negative and disconfirming cases are also important in assisting the categories and emergent (grounded) theory to fit all the data.

Glaser and Strauss (1967, pp. 105–13) suggest that the constant comparison method involves four stages: (i) comparing incidents and data which are applicable to each category; (ii) integrating these categories and their properties; (iii) bounding the theory; (iv) setting out the theory.

The first stage here involves coding of incidents and comparing them with previous incidents in the same and different groups and with other data that are in the same category. For this to happen they suggest *unitizing* – dividing the narrative into the smallest pieces of information or text that are meaningful in themselves, for example, phrases, words, paragraphs. It also involves *categorizing*: bringing together those unitized texts which relate to each other and that can be put into the same category, plus devising rules to describe the properties of these categories and checking that there is

internal consistency within the unitized text contained in those categories.

The second stage involves memoing and further coding. Here the method of constant comparison involves moving beyond comparing one incident with another to comparison of one incident with the properties of the *category* which emerged after comparing incident with incident (Glaser and Strauss, 1967, p. 108).

The third stage – delimitation – occurs at the levels of the theory and the categories (p. 110), in which the major modifications reduce as underlying uniformities and properties are discovered and in which theoretical saturation takes place.

The final stage (writing theory) occurs when the researcher has gathered and generated coded data, memos and a theory which is then written in full.

By going through the previous data, particularly the search for confirming, negative and discrepant cases, the researcher is able to keep a 'running total' of these cases for a particular theory. The researcher also generates alternative theories for the phenomena under investigation (e.g. abduction) and performs the same count of confirming, negative and discrepant cases. Lincoln and Guba (1985, p. 253) argue that the theory with the greatest incidence of confirming cases and the lowest incidence of negative and discrepant cases is the most robust, though this is contestable.

Constant comparison combines the elements of inductive category coding with simultaneously comparing these with the other events and social incidents that have been observed and coded over time and location (LeCompte and Preissle, 1993, p. 256). This enables social phenomena to be compared across categories, where necessary giving rise to new dimensions, codes and categories. Glaser (1978) indicates that constant comparison can proceed from the moment of starting to collect data, seeking key issues and categories, discovering recurrent events or activities in the data that become categories of focus, and expanding the range of categories. This process can continue during the writing-up period, which should be ongoing, so that a model or explanation of the phenomena can emerge which accounts for fundamental social processes and relationships.

The core variable

Through the use of constant comparison a core variable (or core category) is identified: that variable/category which accounts for most of the data and to which as much as possible is related; that variable or category around which most data are focused and to which they relate (Strauss and Corbin, 1990, p. 116). As Flick *et al*. (2004, p. 19) suggest: 'the successive integration of concepts leads to one or more key categories and thereby to the core of the emerging theory'. The core variable is that variable that integrates the greatest number of codes, categories and concepts, and to which most of them are related and with which they are connected. It has the greatest explanatory power; as Glaser (1996) remarks: a concept has to 'earn its way into the theory by pulling its weight' without forcing.

A core variable/category must be central to the category system and the phenomena rather than peripheral to them; it must appear frequently in the data and must fit comfortably and logically to the data rather than be a strained fit. It should have an abstract title but one that is close to the categories and data in question, and it must enable variations to be explained (Strauss and Corbin, 1994).

Saturation

Saturation is reached when no new insights, properties, dimensions, relationships, codes or categories are produced even when new data are added, when all the data are accounted for in the core categories and subcategories and when the coding, categories and data support the emerging theory (Glaser and Strauss, 1967, p. 61; Creswell, 2002, p. 450; Ezzy, 2002, p. 93), and when the variable covers variations and processes (Moghaddam, 2006). Of course one can never know for certain that the categories are saturated, as fresh data may come along that refute the existing theory. The partner of saturation is theoretical completeness, when the theory is able to explain the data fully and satisfactorily.

Theoretical sensitivity

Researchers must possess theoretical sensitivity, i.e. the ability to perceive and notice the important parts of data and to accord them meaning (Strauss and Corbin, 1990, p. 46). It concerns personal qualities in the researcher (p. 41), and sensitivity to the subtleties and complexities of the data and ability to develop theoretical insights into the research. Birks and Mills (2015) comment that it is the researcher's ability to 'recognize and extract from the data' (p. 58) those elements which have relevance and meaning for the emerging theory, without 'forcing' the data into a theory (p. 59). Such sensitivities can be developed from studying relevant literature, professional and personal experience, the processes and procedures followed in the data analysis, continually interacting with the data, reflexivity, standing back from the data to review what is happening, maintaining a critical, perhaps sceptical, attitude to possible explanation, categories and hypotheses concerning the data, i.e. regarding them as provisional only (Strauss and Corbin, 1990, pp. 42–5).

37.5 The strength of the grounded theory

As a consequence of theoretical sampling, coding, constant comparison, the identification of the core variable and the saturation of data, categories and codes, the grounded theory (of whatever is being theorized) emerges from the data in an unforced manner, accounting for all of the data. The adequacy of the derived theory can be evaluated against several criteria. Glaser and Strauss (1967, p. 237) suggest four such main criteria:

1. the closeness of the *fit* between the theory and the data;
2. how readily *understandable* the theory is by laypersons working in the field, i.e. that it makes sense to them;
3. the ability of the theory to apply to a wide range of everyday situations in the same field, i.e. not simply to specific kinds of situation (p. 237);
4. the user of the theory must have sufficient control over their everyday lives so that applying the theory is possible and worthwhile (p. 245).

Strauss and Corbin (1994, pp. 253–6) suggest several criteria for evaluating the theory:

- the adequacy and power of the theory to account for the main concerns of the data;
- the relevance and utility of the theory for the participants;
- the closeness of the fit of the theory to the data and phenomenon being studied, and under what conditions the theory holds true;
- the fit of the axial coding to the categories and codes;
- the ability of the theory to embrace negative and discrepant cases;
- the fit of the theory to literature;
- the appropriateness of the original sample selection, and on what basis;
- what major categories emerged, and what were some of the events, incidents, actions etc. (as indicators) that pointed to some of the major categories?
- on the basis of what categories did theoretical sampling proceed? Was it representative of the categories?
- what were some of the hypotheses pertaining to conceptual relations (that is, among categories), and on what grounds were they formulated and tested?
- were there instances when hypotheses did not hold up against what was actually seen? How were these discrepancies accounted for? How did they affect the hypotheses?
- how and why was the core category selected (sudden, gradual, difficult, easy), and on what grounds?
- were concepts generated and systematically related?
- were there many conceptual linkages between concepts, and were the categories well developed?
- was much variation built into the theory? Were variations explained? Were the broader conditions built into its explanation?
- were change and movement taken into account in the development of the theory?

37.6 Evaluating grounded theory

Strauss and Corbin (1990) indicate that the grounded theory generated should be judged against several criteria (pp. 252–6):

- the reliability, validity and credibility of the data;
- the adequacy of the research process;
- the empirical grounding of the research findings;
- the sampling procedures;
- the major categories that emerged;
- the adequacy of the evidence base for the categories that emerged;
- the adequacy of the basis in the categories that led to the theoretical sampling;
- the formulation and testing of hypotheses and their relationship to the conceptual relations among the categories;
- the adequacy of the way in which discrepant data were handled;
- the adequacy of the basis on which the core category was selected;
- the generation of the concepts;
- the quality of the concepts and the extent to which they are systematically related;
- the number and strength of the linkages between categories, and their conceptual density, leading to their explanatory power;
- the extent of variation that is built into the theory;
- the extent to which the explanations take account of the broader conditions that affected the phenomenon being studied;
- the account taken of emergent processes over time in the research;
- the significance of the theoretical findings.

The emphasis here is on the procedures and not only on the outcomes of the grounded theory research. To this list can be added the criteria of: originality; resonance

(the data, the phenomenon, the participants' experiences and views); usefulness (for different people and groups, for identifying generic processes, for further research, for advancing the field) (Charmaz, 2006, pp. 182–3); 'workability' (practicality and explanatory power); fit with the data; 'relevance' (to the situation, to groups, to researchers, to the field); and 'modifiability' (in light of additional data) (Glaser and Strauss, 1967). Grounded theory is not exempted from the conventional criteria of rigorous research.

37.7 Preparing to work in grounded theory

Glaser (1996) offers some useful practical and personal advice for researchers working with grounded theory. He suggests that researchers must be able to: (a) tolerate uncertainty (as there is no preconceived theory), confusion (cf. Buckley and Waring, 2009, p. 330) and setbacks (e.g. when data disconfirm an emergent theory); (b) avoid premature formulation of the theory, but, by constant comparison, enable the final theory to emerge. They need to be open to what is emerging and not to try to force data to fit a theory but, rather, to ensure that data and theory fit together in an unstrained manner. As he says, 'forcing is a consequence of an inability to handle confusion and regression [feeling stupid] while you study'. Grounded theory, he avers, is an 'experiential methodology', and he advises researchers to 'just do it'! He also indicates that it might not be useful to do much pre-reading since, as he says, 'you never know what you're going to find, so how do you know what to read' (though this is contentious, as discussed earlier and below). He makes the point that, since grounded theory is not easy, the researcher has to be prepared to work hard to be faithful to the rigour of the process.

37.8 Some concerns about grounded theory

There are several concerns raised about grounded theory.

Thomas and James (2006) mount a withering critique of grounded theory, arguing that it 'oversimplifies complex meanings and inter-relationships in data' (p. 768) by focusing on the 'immediately apparent and observable' (p. 769), that it 'constrains analysis' by putting the cart of procedures (theoretical sampling, coding, categorizing, constant comparison, saturation, identification of the core category) before the horse of interpretation, and that it unfairly privileges induction over explanation and prediction (p. 768). They argue that, since grounded theory has many versions, its identity is unclear.

We set out below four further main areas of dispute and criticism with regard to grounded theory.

The meaning and status of theory

Thomas and James (2006) suggest that the term 'theory' is ill-defined and vague in grounded theory, and has many meanings: 'theory' here is 'merely a narrative' rather than an explanation (p. 778), and what grounded theory generates is not a 'theory' at all (p. 780) but simply 'mental constructions', with little explanatory, empirical or predictive power (see also Silverman, 1993, p. 47).

Bryant and Charmaz (2007) comment that grounded theory is epistemologically unclear, makes varied, incommensurate epistemological assumptions (e.g. positivist, inductivist and objectivist versus constructivist, interactionist, deductivist and subjectivist) and is insufficiently related to sociological theory. Waring (2012) notes the contention that there is an unclear separation between discovery and verification. In short, the status and definition of 'theory' in 'grounded theory' is unclear on many fronts. We refer the reader to Chapter 4 for a fuller discussion of 'theory'.

The role of literature and prior disciplinary knowledge

As mentioned earlier, Glaser and Strauss (1967) suggested that the reader should not conduct a literature review in advance of the data analysis, or bring advanced disciplinary knowledge to bear on the analysis, so that the data can speak for themselves, unaffected or contaminated by prior researcher knowledge or preconceptions which might stifle the process of theory generation, and to avoid being '"awed out" by the work of others' (Dunne, 2011, p. 115). Indeed Glaser (1998) argues that, since the researcher does not know in advance what literature will be relevant in the data, conducting a literature review may be time-wasting and inefficient, as it may engage irrelevant material.

This view has attracted much criticism for being artificial (in reality researchers do have some knowledge of the field), unnecessary, disproportionate, unworkable (e.g. for students and researchers who have to provide a literature review as part of a research proposal), over-prescriptive, ideological and, in practice, impossible (e.g. Dunne, 2011; Thornberg, 2012a). The researcher has, and needs to have, prior knowledge in order to understand the field, to be sensitive to the issues and to be able to reflect on the research. Further, as the post-positivists note in Chapter 1, there is no such thing as theory-free observation or a neutral 'God's eye view of the world' (Thornberg, 2012b,

p. 246). Rather, researchers are embedded in a socio-historical, spatio-temporal and ideological world which they cannot simply set aside (cf. Thomas, 2006, pp. 783–4). As social beings, they cannot 'quarantine themselves … from the data they are analysing' (Thomas and James, 2006, p. 781) in their attempts to find an emergent theory.

Silverman (1993, p. 47) suggests that eschewing early literature reviews and disciplinary knowledge fails to acknowledge the implicit theories which guide research from its earliest stages (i.e. data are not theory-neutral but theory saturated), and this should feed into the process of reflexivity in qualitative research.

Dunne (2011) notes that Glaser's insistence on having no advance literature review might be just as time-wasting and inefficient as if one does have one. Literature can provide a rationale for the study, can ensure originality to the study (i.e. that a similar study has not already been conducted and that the study is, indeed, innovative), can avoid the researcher making the same mistakes as other studies and can provide a context for the study (Dunne, 2011; Thornberg, 2012a). It can enable the researcher to be familiar with key concepts ('sensitising concepts'), to spot and circumvent problems (both methodologically and conceptually), to be aware of their own previously unconsciously held preconceptions, and promote clarity in thinking about the concepts and possible theory (Dunne, 2011, p. 116). Indeed literature is a source of inspiration, opening up rather than closing down possibilities and creativity (Thornberg, 2012b, p. 249).

The question of the 'ground' in 'grounded theory'

The grounds – the warrants – for accepting a theory are unclear in grounded theory; are they observation, interpretation, logic, deduction, inference or what? Glaser and Strauss (1967) note that theory is grounded in data and 'discovered', i.e. it has an objective existence in the data that is waiting to be discovered. However, Charmaz (2002), in her constructivist version of grounded theory, suggests that the 'ground' is people, i.e. that the theory is grounded in the meanings and meaning-making that people give to, or construct from, the data, i.e. theory has no objective existence that is waiting to be discovered but is defined by people. Theory is constructed and co-constructed, not discovered, by researchers and participants who bring their own biographies, experiences, contexts and backgrounds to bear on their theories (Charmaz, 2002, 2006; Thornberg, 2012a).

Thomas and James (2006) contend that 'theory' in grounded theory is 'invented' – created – rather than pre-existent and 'discovered' and, anyway, a theory is merely conjectural rather than certain. Indeed they suggest that the fixity and firmness implied in the term 'ground' is inappropriate if one inclines towards Charmaz's view of the 'construction' of a more tenuous, mutable and multiplicity of theory. Hence they suggest replacing the title of the canonical text *The Discovery of Grounded Theory* with *The Invention of Grounded Theory*.

Induction and deduction

Whilst Glaser and Strauss (1967) staunchly advocate induction as the only way of conducting grounded theory and its related data analysis, Strauss and Corbin (1990, 1998) and Corbin and Strauss (2008; 2015) accord a place to deduction and abduction. This marks a major split between Glaser and Strauss. At issue here is that, once deduction and abduction are introduced, the epistemology of grounded theory changes from the generation of emergent theory and hypotheses to the testing and verification of prior existing theory and hypotheses. Glaser adheres to induction, creativity and the imagination in generating theory from the data, whilst Strauss and Corbin allow for hypothesis testing and deductive reasoning, looking *for* data rather than looking *at* data (Robrecht, 1995; Buckley and Waring, 2009, p. 330), which, Glaser (1992, 1998) avers, can lead to 'forcing' data to fit a preconceived theory. Pure induction, argues Thornberg (2012a), is simply impossible.

Generalizability

A concern of grounded theory is how generalizable is the emergent theory. Is it restricted, for example, to being an explanation of the phenomenon in question or does it have wider application, being of a more abstract and lawlike generalization (with the rider that there may not be laws in social science, in contrast to laws in the natural sciences). Does grounded theory aspire to being a 'grand theory' a 'middle-range' theory or an 'empirical theory', as set out in Chapter 4? Is it for the reader to decide whether the theory can apply to a new situation (e.g. Glaser, 1998)? Glaser (1998) argues that a grounded theory must have 'transferability', i.e. must not be bound by the specificities of the particular study in question and must be able to apply to other situations (e.g. through conceptual similarities). This suggests that it is more like the 'middle-range' theory set out in Chapter 4.

Transferability can be, for example, from sample to population and from case to case; the reader may decide how transferable the theory is from one situation to another (Corbin and Strauss, 1990, p. 15). However, this raises questions as to whether a robust theory

should be decided on the grounds of reader judgement alone, though Corbin and Strauss also indicate that theoretical sampling in grounded theory should enable greater generalizability to be achieved. Further, they argue that the generalizability of a grounded theory is achieved, in part through the level of abstraction of the concepts used in the research, particularly in relation to the core category, and in part through the specification of the particular situations to which it might apply.

The dependence on coding

Finally, some versions of grounded theory are heavily dependent on coding, and we refer readers also to the critiques of coding set out in Chapter 34.

Grounded theory, then, though widely used, is not without its challenges (cf. Birks and Mills, 2015). Researchers must decide on its fitness for purpose and, if working with it, must be mindful of its challenges and criticisms.

Companion Website

The companion website to the book provides additional material, data files and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Approaches to quantitative data analysis

CHAPTER 38

Many research data are numerical, and many numerical data are bewildering to researchers. Before moving to specific statistical tests for analysing data, we introduce some important foundational concepts in this chapter. These include:

- scales of data
- parametric and non-parametric data
- descriptive and inferential statistics
- kinds of variables
- hypotheses
- one-tailed and two-tailed tests
- distributions

38.1 Introduction

The prospect of analysing numerical data sends shivers down the spines of many novice researchers who not only baulk at the thought of statistics but hold fundamental objections to what they see as 'the mathematization of nature' (Horkheimer, 1972). Most concepts in education, some assert, are simply not reducible to numerical analysis. Statistics, they say, combine refinement of process with crudity of concept. We do not hold with any of this. Quantitative data analysis has no greater or lesser importance than qualitative analysis and *vice versa*. Its use is entirely dependent on *fitness for purpose*. Arbitrary dismissal of numerical analysis is mere ideology or prejudice.

Quantitative data analysis is a powerful research form. It is often associated with large-scale research, but can also serve smaller-scale investigations, with case studies, action research, correlational research and experiments. In the following chapters we show how numerical data can be reported and introduce some of the most widely used statistics that can be employed in their analysis.

Numerical analysis can be performed using software, for example, the Statistical Package for Social Sciences (SPSS), Minitab, Excel, SAS, Statistica, R. Software packages apply statistical formulae and carry out computations. With this in mind, we avoid extended outlines of statistical formulae, though we do provide details where considered useful. Our primary aim is to explain the concepts that underpin statistical analyses and to do this in as user-friendly a way as possible. Lest our approach should raise purist eyebrows, we provide greater detail in references and we signal these where appropriate. Our outline commentary is closely linked to SPSS, a widely used statistical package for social sciences. An introductory SPSS manual to this volume is provided in an accompanying website (including print-outs of data analysis), together with comments on what they show. It is often the case that such outputs can clarify issues more straightforwardly than extended prose. We also include a guide to all the SPSS files held on the companion website to this chapter.

In this chapter we identify some key concepts in numerical analysis (scales of data, parametric and non-parametric data, descriptive and inferential statistics, dependent and independent variables). Throughout this chapter and subsequent chapters we indicate how to report analysis; these are collected together in a single file on the accompanying website. Material in the accompanying web site also refers to statistical tables, which can also be found on the website.

In this chapter we introduce some basic concepts and terms, which are taken up in later chapters on descriptive statistics and inferential statistics. Bearing in mind the range of statistics covered, Chapter 44 maps out key statistics that are available to the researcher.

38.2 Scales of data

Before one can advance very far in the field of data analysis one needs to distinguish the kinds of numbers with which one is dealing. This takes us to the commonly reported issue of scales or levels of data, and four are identified, each of which, in the order given below, subsumes its predecessor.

The *nominal* scale simply denotes categories, 1 means such-and-such a category, 2 means another and so on, for example, '1' might denote males, '2' might denote females. The categories are mutually exclusive and have no numerical meaning. For example, consider

numbers on a football shirt: we cannot say that the player wearing number 4 is twice as anything as a player wearing a number 2, nor half as anything as a player wearing a number 8; the number 4 simply identifies a category, and indeed nominal data are frequently termed categorical data. The data classify, but have no order. Nominal data include items such as sex, age group (e.g. 30–35, 36–40), subject taught, type of school, socio-economic status. Nominal data denote discrete variables, entirely separate categories, for example, according females the number 1 category and males the number 2 category (there cannot be a 1.25 or a 1.99 position). The figure is simply a conveniently short label.

The *ordinal* scale classifies and also introduces an order into the data. These might be rating scales where, for example, 'strongly agree' is stronger than 'agree', or 'a very great deal' is stronger than 'very little'. It is possible to place items in an order, weakest to strongest, smallest to biggest, lowest to highest, least to most and so on, but there is still an absence of a standard metric – a measure using calibrated or equal intervals. Therefore one cannot assume that the distance between each point of the scale is equal, i.e. the distance between 'very little' and 'a little' may not be the same as the distance between 'a lot' and 'a very great deal' on a rating scale. One could not say, for example, that, in a five-point rating scale (1=strongly disagree; 2=disagree; 3=neither agree nor disagree; 4=agree; 5=strongly agree), point 4 is in twice as much agreement as point 2, or that point 1 is in five times more disagreement than point 5. However, one could place them in an order: 'very little', 'a little', 'somewhat', 'a lot', 'a very great deal', or 'strongly disagree', 'disagree', 'neither agree nor disagree', 'agree', 'strongly agree', i.e. it is possible to rank the data according to rules of 'lesser than' or 'greater than', in relation to whatever value is included on the rating scale. Ordinal data include items such as rating scales and Likert scales, and are frequently used in asking for opinions and attitudes.

The *interval* scale introduces a metric – a regular and equal interval between each data point – as well as keeping the features of the previous two scales, classification and order. This lets us know 'precisely how far apart are the individuals, the objects or the events that form the focus of our inquiry' (Cohen and Holliday, 1996, p. 9). As there is an exact and same interval between each data point, interval level data are sometimes called *equal-interval scales* (e.g. the distance between 3 degrees Celsius and 4 degrees Celsius is the same as the distance between 98 degrees Celsius and 99 degrees Celsius). However, in interval data, there is no true zero. Let us give two commonly cited examples. In Fahrenheit degrees, the freezing point of water is 32 degrees, not zero, so we cannot say, for example, that 100 degrees Fahrenheit is twice as hot as 50 degrees Fahrenheit, because the measurement of Fahrenheit did not start at zero. In fact, twice as hot as 50 degrees Fahrenheit is 68 degrees Fahrenheit (({50–32}×2)+32). Let us give another example. Many IQ tests commence their scoring at point 70, i.e. the lowest score possible is 70. We cannot say that a person with an IQ of 150 has twice the measured intelligence as a person with an IQ of 75 because the starting point is 70; a person with an IQ of 150 has twice the measured intelligence as a person with an IQ of 110, as one has to subtract the initial starting point of 70 ({150–70}/2). In practice, the interval scale is rarely used, and the statistics that one can use with this scale are, to all intents and purposes, the same as for the fourth scale: the ratio scale. In fact, some statistical software packages (e.g. SPSS) combine interval and ratio scales in a single type; in SPSS this is called 'scale'.

The *ratio* scale embraces the main features of the previous three scales – classification, order and an equal-interval metric – but adds a fourth, powerful feature: a true zero. This enables the researcher to determine proportions easily – 'twice as many as', 'half as many as', 'three times the amount of', and so on. Because there is an absolute zero, all of the arithmetical processes of addition, subtraction, multiplication and division are possible. Measures of distance, money in the bank, population, time spent on homework, years teaching, income, marks on a test and so on are all ratio measures as they are capable of having a 'true' zero quantity. If I have $1,000 in the bank then it is twice as much as if I had $500 in the bank; if I score 90 per cent in an examination then it is twice as many as if I had scored 45 per cent. The opportunity to use ratios and all four arithmetical processes renders this the most powerful level of data. Interval and ratio data are continuous variables that can take on any value within a particular, given range. Interval and ratio data typically use more powerful statistics than nominal and ordinal data.

The delineation of these four scales of data is important, as much hinges on this. The consideration of which statistical test to use depends on the scale of data; it is incorrect to apply many statistics which can only be used at interval or ratio scales of data to nominal or ordinal data. For example, one should not apply averages (means) to nominal data, nor use t-tests and Analysis of Variance (discussed later) to ordinal data. Which statistical tests can be used with which data are set out clearly in subsequent chapters.

To close this section we record Wright's (2003, p. 127) view that the scale of measurement is not always inherent to a particular variable, but something that researchers 'bestow on it based on our theories of that variable'. For example, are the five points in an ordinal scale (e.g. 'very little', 'a little', 'somewhat', 'a lot', 'a very great deal') really better described as categorical, and are the categories in the ordinal scale (e.g. 'strongly disagree', 'disagree', 'neither agree nor disagree', 'agree', 'strongly agree') more fittingly described as categorical, two of them being categories of disagreement, two being categories of agreement and one being neutrality? What is being suggested here is that we have to justify classifying a variable as nominal, ordinal, interval or ratio, and not just assume that it is self-evident. Much of the time this is beyond dispute, but ordinal data may be problematic here.

38.3 Parametric and non-parametric data

Non-parametric data are those which make no assumptions about the population, usually because the characteristics (numerical parameters) of the population are unknown. Parametric data assume knowledge of the characteristics of the population, in order for inferences to be able to be made securely; inferential statistics are premised on a normal, Gaussian curve of distribution, as, for example, in reading scores, in order to be able to generalize to the wider population (though Wright (2003, p. 128) suggests that normal distributions are actually rare). In practice this distinction means the following: nominal and ordinal data are often considered to be non-parametric, whilst interval and ratio data are often considered to be parametric data (unless, for example, the data are skewed). The distinction is important, as, for the four scales of data, the consideration of which statistical test to use is dependent on the kinds of data: it is often incorrect to apply parametric statistics to non-parametric data, though it is possible to apply non-parametric statistics to parametric data if those data do not conform to the curve of distribution, being skewed or unevenly distributed. Statistics for parametric data tend to be more powerful than those for non-parametric data, though such power is bought at the price of, for example, conformity to the normal curve of distribution and random samples. Non-parametric data are often derived from questionnaires and surveys (though these can also include parametric data), whilst parametric data tend to be derived from experiments and tests (e.g. examination scores).

38.4 Descriptive and inferential statistics

Descriptive statistics do exactly what they say: they describe and present data, for example, in terms of summary frequencies. No attempt is made to infer or predict population parameters, and they are concerned simply with enumeration and organization. This includes:

- the mode (the score obtained by the greatest number of people);
- the mean (the average score);
- the median (the score obtained by the middle person in a ranked group of people, i.e. it has an equal number of scores above it and below it);
- minimum and maximum scores;
- the range (the distance between the highest and the lowest scores);
- the variance (a measure of how far scores are from the mean, calculated as the average of the squared deviations of individual scores from the mean);
- the standard deviation (a measure of the dispersal or range of scores, calculated as the square root of the variance, yielding the average of all the individual deviations of scores from the mean);
- the standard error (the standard deviation of sample means);
- the skewness (how far the data are asymmetrical in relation to a 'normal' curve of distribution);
- kurtosis (how steep or flat is the shape of a graph or distribution of data; a measure of how peaked a distribution is and how steep is the slope or spread of data around the peak).

Such statistics make no inferences or predictions; they simply report what has been found, in a variety of ways.

Inferential statistics, by contrast, strive to make inferences and predictions based on the data gathered. They infer or predict population parameters or outcomes from simple measures, for example, from sampling and from statistical techniques, and they use information from a sample to reach conclusions about a population, based on probability. Such statistics include hypothesis testing, regression and multiple regression, difference testing (e.g. t-tests and Analysis of Variance), factor analysis and structural equation modelling. Sometimes simple frequencies and descriptive statistics may speak for themselves, and the careful portrayal of descriptive data may be important. However, often it is inferential statistics that are more valuable for researchers, and typically these are more powerful.

38.5 Kinds of variables

A variable is a condition, factor or quality that, as its name suggests, can vary, for example, in quantity, intensity etc., from one case to another; it is the opposite of a constant, which does not vary between cases.

Dependent and independent variables

Research often concerns relationships between variables (a variable can be considered as a construct, operationalized construct or particular property in which the researcher is interested). An independent variable is, as its name suggests, a variable which is not affected by another variable, it is independent of other variables. Typically it is a variable that the researcher can manipulate or control; it is often considered to be a stimulus that influences a response, an antecedent which may be manipulated or modified (e.g. under experimental or other conditions that might control the amount or frequency of something) to affect an outcome. A dependent variable, on the other hand, is, as its name suggests, a variable whose (numerical) value depends to some degree on that of one or more independent variables. This is a fundamental concept in many statistics and it is the basis of a lot of statistical modelling.

For example, we may wish to see if doing more homework (independent variable) increases students' performance in, say, mathematics (dependent variable). We increase the homework and measure the result, and we notice, for example, that the performance increases on the mathematics test. The independent variable has produced a measured outcome. Or has it? Maybe: (a) the threat of the mathematics test increased the students' concentration, motivation and diligence in class; (b) the students liked mathematics and the mathematics teacher, and this caused them to work harder, not the mathematics test itself; (c) the students had a good night's sleep before the mathematics test and, hence, were refreshed and alert; (d) the students' anticipated performance in the mathematics test, in fact, influenced how much homework they did – the higher the anticipated marks, the more they were motivated to doing mathematics homework; (e) the increase in homework increased the students' motivation for mathematics and this, in turn may have caused the increased performance in the mathematics test; (f) the students were told that if they did not perform well on the test then they would be punished, in proportion to how poorly they scored.

What one can observe here is important. In respect of (a) there are other *extraneous* variables which must be factored into the causal relationship (i.e. in addition to the homework). In respect of (b) the assumed relationship is not really present; behind the coincidence of the rise in homework and the rise in the test result is a stronger causal relationship of the liking of the subject and the teacher which caused the students to work hard, a by-product of which was the rise in test scores. In respect of (c) an *intervening* variable was at work (a variable which affected the process of the test but which was not directly observed, measured or manipulated). In respect of (d), in fact the anticipated test result caused the increase in homework, and not *vice versa*, i.e. the direction of causality was reversed. In respect of (f), the amount of increase was negatively correlated with the amount of punishment: the greater the mark, the lesser the punishment. In fact, what may be happening here is that causality may be less in a linear model and more multi-directional and multi-related, more like a web than a line (cf. Morrison, 2009, 2012).

This example indicates a range of issues in the discussion of dependent and independent variables:

- the direction of causality is not always clear (an independent variable may, in turn, become a dependent variable and *vice versa*);
- causality may be bi-directional or multi-directional;
- assumptions of association may not be assumptions of causality;
- there may be a range of other factors which have a bearing on a dependent variable;
- there may be causes (independent variables) behind the identified causes (independent variables) that have a bearing on the dependent variable;
- the independent variable may cause something else, and it is the something else that causes the outcome (dependent variable);
- causality may be non-linear rather than linear;
- the direction of the relationship may be negative rather than positive;
- the strength/magnitude of the relationship may be unclear.

Many statistics operate with dependent and independent variables (e.g. experiments using t-tests and Analysis of Variance, regression and multiple regression); others do not (e.g. correlational statistics, factor analysis). If one uses tests which require independent and dependent variables, caution has to be exercised in assuming which is or is not the dependent or independent variable, and whether causality is as simple as the test assumes. Further, many statistical tests are based on linear relationships (e.g. correlation, regression and multiple regression, factor analysis) when, in fact the relationships may not be linear (some software programs, e.g. SPSS, have the capability for handling non-linear relationships). The researcher has to make

a fundamental decision about whether, in fact, the relationships are linear or non-linear, and select the appropriate statistical tests with these considerations in mind.

Moderator and mediator variables

In addition to independent and dependent variables, there are also moderator and mediator variables. In conducting a correlation one might calculate the correlation coefficient between two variables, say hours of study and performance in a test of mathematics, to be 0.95, i.e. very strong indeed, but, when a third variable is introduced (as in a partial correlation, see Chapter 40), say motivation level for mathematics, the correlation coefficient drops to 0.12, a very weak correlation; in other words, the third variable (motivation for mathematics) is exerting a strong moderation effect.

A moderator variable is one which affects the strength and/or the direction of a relationship between two other variables, for example, between an independent and a dependent variable, i.e. whose values influence the values of another variable. For example, school leadership might moderate the relationship between teacher commitment and school effectiveness: a highly effective leader might increase teacher commitment and, hence, school effectiveness, while a weaker leader might reduce teacher commitment and, hence, reduce school effectiveness.

Moderators are synonymous with interactions. For example, in Analysis of Variance, a moderating variable, say hours of part-time study, might affect the interaction between stress level and performance in the history test, i.e. the interaction of stress level and hours of study might exert an effect on the dependent variable (performance in the history test).

An example of a moderator variable is thus: socio-economic status (A) might have a relationship to performance in the international mathematics test (C), but this might be affected by the age of the student (moderator variable B), so, for example, (A) may have only a small effect on performance in the international mathematics test (C) for a student aged 11 but a larger effect on a student aged 15, and a much larger effect on a student aged 17. Indeed, Bourdieu's work on cultural capital (1976) suggests that this is the case. Here the moderator variable B appears to have an influence on the strength of the relationship between A and C.

A mediator variable is one which explains the relationship between an independent and dependent variable, or between two other variables (Baron and Kenny, 1986). A mediator variable (B) receives the effect of one independent variable (A) and this affects the outcome variable (C). Here the relationship between A and C is indirect because it is mediated by B; it goes through B.

For example, consider the relationship between (A) socio-economic status and (C) performance on, say, an international test of mathematics. Here hours of study (B) may be a mediating variable as it explains the relationship between A and C: socio-economic status (A) affects hours of study (B) which affects performance in the international mathematics test (C). Socio-economic status (A), here, has an indirect relationship with performance in the international mathematics test (C) as it goes through variable (B); B is a conduit which renders A an indirect independent variable: A→B→C.

Moderator and mediating variables can be explored through controlling for variables, for example in partial correlations and structural equation modelling. Moderator and mediator variable are intimately connected to causal modelling, and we refer the reader to Chapter 6.

To draw these points together, the researcher will need to consider:

- What scales of data are there?
- Are the data parametric or non-parametric?
- Are descriptive or inferential statistics required?
- Do dependent and independent variables need to be identified?
- Do the research and data analysis need to take account of moderating and mediating variables?
- Are the relationships considered to be linear or non-linear?

The prepared researcher will need to consider the mode of data analysis that will be employed. This is very important as it has a specific bearing on the form of the instrumentation used. For example, a researcher will need to plan the layout and structure of a questionnaire survey very carefully in order to assist data entry for computer reading and analysis; an inappropriate layout may obstruct data entry and subsequent computer processing. The planning of data analysis will need to consider:

- what needs to be done with the data when they have been collected – how the data will be processed and analysed;
- how the results of the analysis will be verified, cross-checked and validated.

Decisions will need to be taken with regard to which statistical tests to use in data analysis, as this affects the layout of items (e.g. in a questionnaire) and the computer packages that are available for processing

quantitative and qualitative data, for example, SPSS and NVivo respectively.

Categorical, discrete and continuous variables

A categorical variable is a variable which has categories of values. For example, the variable 'sex' has two values: male and female; it is a dichotomous variable. In a rural community with, say, four local schools, the variable 'school attended' will have four values, one for each school. If we are looking at the types of food in school meals we may want to have three categories: carbohydrates, proteins and fats; each of these is a category of the variable 'food'.

A discrete variable has a finite number of values of the same item, with no fractions of the value (e.g. the number of illnesses a person has had, the number of mealtimes a person has each day). Here there are no fractions of a value – a person cannot have half an illness or half a mealtime; they either have the illness or not, they either have the mealtime or not.

A continuous variable, as its name suggests, can vary in quantity, for example, money in the bank, monthly earnings, numbers of students present in a class. Here there are equal intervals, and, for ratio data, a zero (it is possible to have no money in the bank, or to have no earnings, or for a class of students to have none present that day).

Categorical variables yield categorical data. Continuous variables yield interval and ratio data (though in SPSS these are combined in the classification of 'scale' data). Depending on the kind of variable one has will be the kinds of statistics that can be used. This is addressed in subsequent chapters.

Kinds of analysis

Univariate analysis examines differences among cases within one variable. Bivariate analysis looks for a relationship between two variables. Multivariate analysis looks for a relationship between two or more variables. Different statistics are used, depending on whether one is working with univariate, bivariate or multivariate analysis.

Hence, in approaching statistical processing and analysis, the researcher will need to decide:

- the scales of the data being used (categorical, ordinal, interval, ratio);
- the kind of data being used (parametric, non-parametric);
- the kinds of variables being used (categorical, discrete, continuous, independent, dependent, moderator, mediator);
- the kinds of statistics to be used (descriptive, inferential).

38.6 Hypotheses

Research in a hypothetico-deductive mode and research that uses statistics often commence with one or more hypotheses. This is the essence of hypothesis testing in quantitative research. Typically hypotheses fall into different types. The *null hypothesis*, a major type of hypothesis, states that, for example, there is *no* relationship between two variables, or that there has been *no* change in participants between a pre-test and a post-test, or that there is *no* difference between three school districts in respect of their examination results, or that there is *no* difference between the voting of males and females on such-and-such a factor. Null hypothesis significance testing (NHST) is addressed in Chapter 39.

The point here is that by casting the hypothesis in a null form, the burden of proof is placed on the researcher *not* to support that null hypothesis. The task is akin to a jury starting with a presumption of innocence and having to prove guilt beyond reasonable doubt. Not only is it often easier simply to support a straightforward positive hypothesis, but, more seriously, even if that positive hypothesis is supported, there may be insufficient grounds for accepting that hypothesis, as the finding may be consistent with other hypotheses. For example, let us imagine that our hypothesis is that a coin is weighted and, therefore, unfair. We flip the coin 100 times, and find that 60 times out of 100 it comes out as heads. It would be easy to jump to the conclusion that the coin is weighted, but, equally easily, other reasons may account for the result. Of course, if the coin were to come out as heads 99 times out of 100 then perhaps there would be greater truth in the hypothesis. Null hypothesis testing is a stronger version of evidence, requiring not only that the negative hypothesis be 'not supported', but also indicating a cut-off point only above which the null hypothesis is 'not supported', and below which the null hypothesis is supported. In our coin example, it may be required to find that heads comes up 95 times out of 100, or 99 times out of 100, or even 999 times out of 1,000, to say, with increasing confidence in respect of these three sets of figures, that the null hypothesis is not supported. We discuss this in terms of statistical significance in Chapter 39.

We use terminology carefully here. Some researchers state that the null hypothesis is 'rejected'; others say that it is 'confirmed' or 'not confirmed'; others say that it is 'accepted' or 'not accepted'. We prefer the

terminology of 'supported' or 'not supported'. This is not mere semantics or pedantry; rather it signals caution. Rejecting a null hypothesis is not the same as 'not confirming' or 'not supporting' that null hypothesis, as rejection implies an absolute and universal state which the research will probably not be able to demonstrate, being bounded within strict parameters and not being applicable to all cases. Further, 'confirming' and 'not confirming', like 'rejecting', is too strong, absolute and universal a set of terms for what is, after all, research that is bounded and within delineated boundaries. Similarly, one cannot 'accept' a null hypothesis as a null hypothesis can seldom be proved unequivocally (though there are occasions when it can, e.g. if the means and standard deviations between, say, the maths test scores of two groups are equal).

A second type of hypothesis is termed the *alternative hypothesis*. Whereas the null hypothesis states that there is *no* such-and-such (e.g. change, relationship, difference), the alternative hypothesis states that there *is* such-and-such, for example: there *is* a change in behaviour of the school students; there *is* a difference between students' scores on mathematics and science; there *is* a difference between the examination results of five school districts; there *is* a difference between the pre-test and post-test results of such-and-such a class. This kind of hypothesis is often supported when the null hypothesis is 'not supported', i.e. if the null hypothesis is not supported then the alternative hypothesis is.

The two kinds of hypothesis are usually written thus:

H_0: the null hypothesis
H_1: the alternative hypothesis

Sometimes the alternative hypothesis is written as H_A. So, for example, the researcher could write null hypotheses and alternative hypotheses thus:

H_0: There is no statistically significant difference between the results of the control group and experimental group in the post-test of mathematics
or There is no statistically significant difference between males and females in the results of the English examination
or There is no statistically significant correlation between the importance given to a subject and the amount of support given to it by the headteacher

H_1: There is a statistically significant difference between the control group and experimental group in the post-test of mathematics
or There is a statistically significant difference between males and females in the results of the English examination
or There is a statistically significant positive correlation between examination scores in mathematics and science

(We address statistical significance in Chapter 39.) The null hypothesis requires rigorous evidence *not* to support it. The alternative hypothesis is taken up when the first – null – hypothesis is not supported. The latter is the logical opposite of the former. One commences with the former and casts the research in the form of a null hypothesis, only turning to the latter if it is found that the null hypothesis is not supported.

A hypothesis can be directional or non-directional. A directional hypothesis states the kind of difference or relationship between two conditions or two groups of participants. For example:

- Students who do homework without the television switched on in their room whilst working produce *better* results than those who do homework with the television switched on.
- Students who have a computer at home do better in exams than people who do not.
- People remember the words that appear early in a list better than the words that appear later.
- People who are given a list of emotionally charged words recall more than participants given a list of neutral words.

Here one can see the direction of the hypothesis ('better', 'more than').

By contrast, a non-directional hypothesis simply predicts that there will be a difference or relationship between two conditions or two groups of participants, but it does not state the direction of the difference (e.g. 'more than', 'less than, 'better than', 'worse than'), for example:

- Students who do homework without the television switched on in their room whilst working produce different results from those who do homework with the television switched on.
- Students who have a computer at home perform differently in exams than people who do not.
- People remember a different number of words that appear early in a list than the words that appear later.

- People who are given a list of emotionally charged words recall a different number than participants given a list of neutral words.

Here there is a difference, but the *direction* of that difference is not made explicit. The stronger of these two types of hypothesis is the directional hypothesis because it makes a stronger claim than the non-directional hypothesis. In hypothesis testing the researcher:

1. formulates a hypothesis;
2. measures the variables involved and examines the relationship between them;
3. calculates the probability of obtaining such a relationship if there were no relationship by chance, i.e. if the null hypothesis is true. If the calculated probability is small enough, it suggests that the pattern of findings is unlikely to have arisen by chance, and probably reflects a genuine relationship.

38.7 One-tailed and two-tailed tests

In using statistics, researchers are sometimes confronted with the decision of whether to use a one-tailed or a two-tailed test. Which to use is a function of the kind of result one might predict. In a one-tailed test one predicts, for example, that one group will score more highly than the other, whereas in a two-tailed test one makes no such prediction. The one-tailed test is a stronger test than the two-tailed test as it makes assumptions about the population and the direction of the outcome (i.e. that one group will score more highly than another), and hence, if supported, is more powerful than a two-tailed test. A one-tailed test is used with a directional hypothesis (e.g. 'students who do homework *without* the TV on produce *better* results than those who do homework with the TV playing'). A two-tailed test is used with a non-directional hypothesis (e.g. 'there is a *difference* between homework done in noisy or silent conditions'). Here the directional hypothesis indicates 'more' or 'less', whereas the non-directional hypothesis indicates only difference, and not where the difference may lie.

For example, let us imagine that we run a 'true' experiment to see if students who do homework *without* the TV on produce *better* results than those who do homework with the TV playing. The results are shown in Figure 38.1.

We can see here that there is an overlap between the two sets of scores, but that the scores of the group which did not have the television switched on whilst doing homework are much higher than the scores of the group whose television is switched on whilst doing homework. Our directional hypothesis is supported, and we have used a one-tailed test to test the hypothesis.

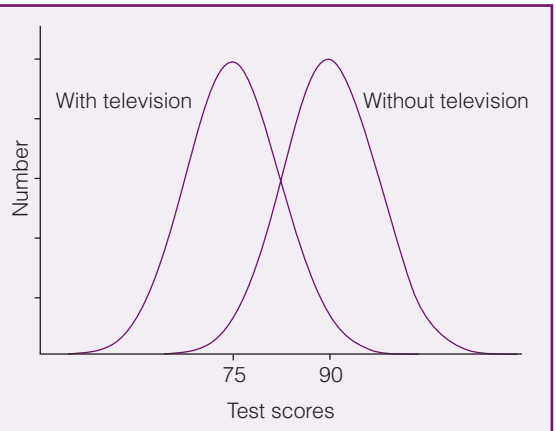


FIGURE 38.1 Test scores of two groups

In graphical terms, we can portray the results of a one-tailed test that predicts high scores in Figure 38.2.

Here the prediction is that those students who work without the television switched on score more highly (the shaded area at one end (tail) of the graph). The '5%' indicates that we are predicting with a 95 per cent degree of certainty that the results will be higher.

By contrast, we can portray the results of a one-tailed test that predicts low scores in Figure 38.3.

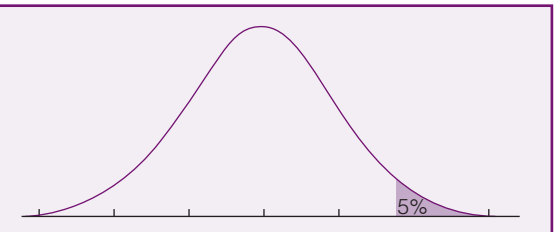


FIGURE 38.2 The predictions of a one-tailed test that predicts a higher score

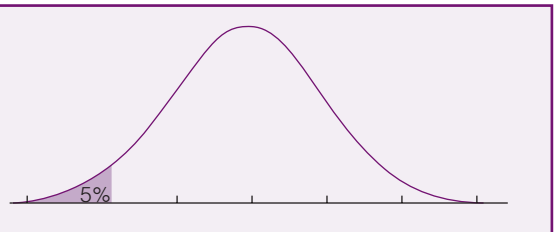


FIGURE 38.3 The predictions of a one-tailed test that predicts a lower score

Here the prediction is that those students who work with the television switched on score lower (the shaded area at one end (tail) of the graph). Again, because we predict the direction of the result, we use a one-tailed test. Here the '5%' indicates that we are predicting with a 95 per cent degree of certainty that the results will be lower.

Figure 38.4 indicates the results for a two-tailed test. Here we can see that there are two shaded areas, one at each end (tail) of the graph. Because we have not predicted the direction of the result (it could be higher or lower), the burden of proof is higher, i.e. instead of having a 95 per cent certainty (the 5 per cent of the two previous figures), we distribute that 5 per cent between the two tails, each of which is 2.5 per cent, i.e. we need to demonstrate a 97.5 per cent certainty level in the result.

38.8 Confidence intervals

In working with statistics, researchers need to know the confidence that they can have in the accuracy of their findings. Indeed the American Psychological Association (2010, p. 34) requires confidence intervals to be included in publications of research findings. A confidence interval is a range of values for the results obtained, and researchers need to know how confident they can be that their particular result falls within that acceptable range, i.e. that the range will include the result in question. For example, a researcher may say 'my survey finding, that 15 per cent of students prefer to have a female teacher of mathematics, is true, give or take 5 per cent', i.e. the percentage range could be from 10 to 20 per cent (15–5 and 15+5). For example, if the researcher had repeated her survey with a different group, her result might be that 17 per cent of students preferred to have a female teacher of mathematics; this still falls within the acceptable range of 10–20 per cent. Confidence intervals indicate that the acceptable range, and how much confidence can be placed in suggesting that this range, includes the particular finding in question.

FIGURE 38.4 The predictions of a two-tailed test

Researchers may know, for example, that their finding of an average score from a sample is an approximation of the likely average score in the population, but they need to know how good their approximation is (Dancey and Reidy, 2011, p. 109). Here confidence intervals feature, as they indicate the range of possible scores in the population, and the confidence that this range will include the researcher's score.

Many statistics packages have a default setting of returning a 95 per cent confidence interval. Confidence intervals are important in inferential statistics, as they enable the researcher to state how confident they can be in inferring that the score found will be in the range (the interval) of acceptable scores for that result.

Say, for example, a researcher wishes to calculate whether the difference between the means of two test results (maths and history) is reliable. She conducts a t-test (see Chapter 41) and finds that the average difference in scores between the maths test and the history test is 0.350, with a standard error of the difference being 0.464 (here we do not go into how these scores are computed, as this requires knowledge of the t-test, and we address this in Chapter 41; for the present we ask the reader simply to accept these data). Now the computer software (in this case SPSS) shows that a 95 per cent confidence interval indicates that the range of an average difference should be between –0.579 and +1.278. In other words, the researchers could be 95 per cent confident that the score she found for the average difference (0.350) should lie between –0.579 and +1.278. In the instance here, this was found.

Confidence intervals can be used with descriptive and inferential statistics, and effect size calculations (cf. Ellis, 2010, p. 19). They are affected by sample size; Torgerson and Torgerson (2008, p. 135) note that small sample sizes tend to have wider confidence intervals, suggesting, conversely, that a large sample will reduce the range of the confidence interval, i.e. a large sample will give greater precision to the result and the researcher can have 95 per cent certainty of her result being included in that narrower range. A confidence interval, as its name suggests, sets out the range (the interval) of likely results, and the often-used 95 per cent confidence interval is a measure of certainty, i.e. that the researcher can have a 95 per cent confidence that the range of possible results here (the interval) will include the result that she found for her research.

38.9 Distributions

In everyday life many variables tend to be normally distributed. For example, we may say that most men are about such-and-such an average height. Some of

course will be taller, and some shorter; there is a range. A smaller number will be *much* taller or *much* shorter; an even smaller number will be *very much* taller or *very much* shorter; a very small number will be *extremely* tall or *extremely* short. We can present the results on a normal curve of distribution (Figure 38.5).

Many statistics (e.g. inferential statistics) assume a normal curve of distribution and, indeed, researchers should test for the nature of the distributions to see if they conform to the normal curve, as this has an effect on the choice of statistics. For example, if the distribution of the data conforms to the normal curve then this might enable inferential parametric statistics to be calculated, whereas if the distribution does not conform to the normal curve, even if the data are interval or ratio, then this might require the use of distribution-free, i.e. non-parametric, statistics.

The normal curve of distribution is a smooth, perfectly symmetrical (bell-shaped) curve; it is symmetrical about the inflection point, with the mean (the average score) at the point of inflection, and its tails are assumed to meet the x-axis at infinity. Here 68.3 per cent of people will fall within one standard deviation of the mean (a measure of the average variance from the mean). In our example we might assume that the majority of men (68.3 per cent) will be either just about or just below the mean height. Then, if we look at Figure 38.5, we can see that a smaller proportion (95.4 per cent minus 68.3 per cent = 27.1 per cent) are *much* taller or *much* shorter (between one standard deviation and two standard deviations away from the mean), and an even smaller proportion (99.7 per cent minus 95.4 per cent = 4.3 per cent) are *very* tall or *very* short, (even further away from the mean (between two and three standard deviations)), and only a very tiny proportion (100 per cent minus 99.7 per cent = 0.3 per cent) are *extremely* tall or *extremely* short (more than three standard deviations away from the mean).

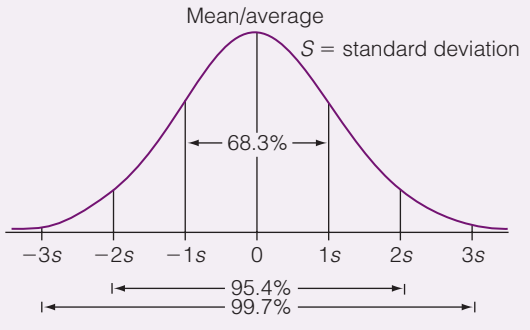


FIGURE 38.5 The normal curve of distribution

In educational research that uses statistics, many statistical calculations assume that the population is distributed normally and then compare the data collected from the sample to the population, allowing inferences to be made about the population (e.g. in random sampling). The assumption of the normal curve of distribution enables researchers to measure all normal distributions of a variable, regardless of the units in which that variable is initially measured, and to be able to generalize to a wider population.

Of course this is not always the case; rarely, if ever, in real life are data distributed so neatly. Rather than being symmetrical, data might be skewed in different ways. They may be positively or negatively skewed, as in Figure 38.6.

Skewed distributions are not symmetrical; a positively skewed distribution has the tail skewed to the

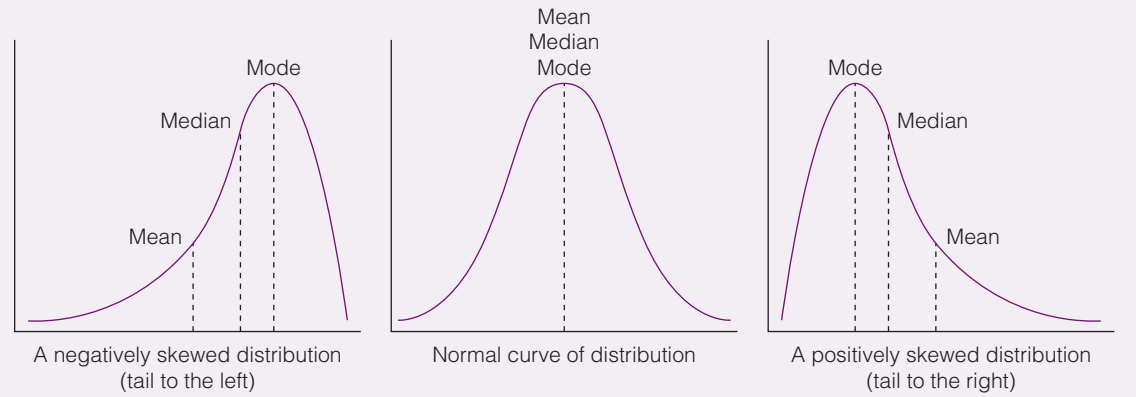


FIGURE 38.6 Skewed distributions

right, whilst a negatively skewed distribution has the tail skewed to the left. This has important implications for even the simplest statistics calculated. For example, whilst the mean (average) may be useful for normal distributions, in skewed distributions it is an unreliable measure, as it is affected by the long tail (positive or negative). Similarly the mode (the particular score registered by the most voters) may not be an accurate measure of the distributions. In both of these cases, the median score (the score which is given by the middle person, e.g. in a test) is more reliable.

An example of skewness is given in Figure 38.7. The figure presents a line graph to show how respondents voted on how well learners are guided and supported in their learning, awarding marks out of ten for the voting, with a sample size of 400 respondents.

Here the data are skewed, with more votes being received at the top end of the scale. There is a long tail going to the negative end (left-hand side) of the scores, so, even though the highest scores are given at the top end of the scale, we say that this graph has a negative skew because there is a long tail down.

By contrast, let us look at a graph of how much staff voluntarily take on roles in the school, with 150 votes received and awarding marks out of ten (Figure 38.8).

Here a long tail goes towards the upper end of the scores, and the bulk of the scores are in the lower range. Even though most of the scores are in the lower range, because the long tail is towards the upper end (right-hand side) of the scale this is termed a positive skew.

Further, a graph of distribution may not always have the same bell-shaped features of the normal curve. For example it may be flatter than normal (platykurtic) or steeper than normal (leptokurtic) (see Figure 38.9).

The measure of steepness of the curve is termed 'kurtosis'; many statistics packages calculate the measure of kurtosis. Normal distributions have a skewness of zero and a measure of kurtosis of zero: a platykurtic distribution has a negative value of kurtosis, whilst a leptokurtic distribution has a positive value of kurtosis. The degree of kurtosis may affect the reliability of the statistics that are used or the inferences that are made from them. For example, a platykurtic distribution may not have a problem with outliers, whilst a leptokurtic distribution may; further, many statistics assume normal kurtosis, rather than unduly flat or steep kurtosis.

The skewness and kurtosis of the data are important features to observe in data, and to which to draw attention. Researchers must check to see if the distributions of their data conform to the normal curve. There are different ways of doing this. One way is simply to construct a histogram and look at its overall shape to see if

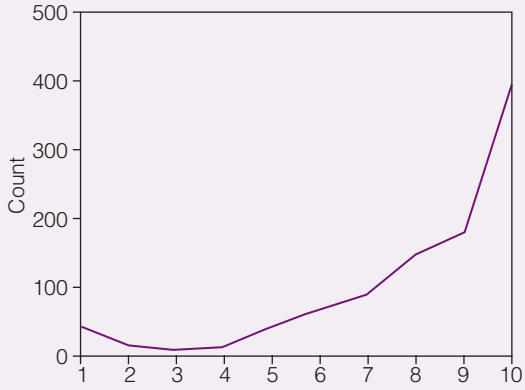


FIGURE 38.7 How well learners are cared for, guided and supported

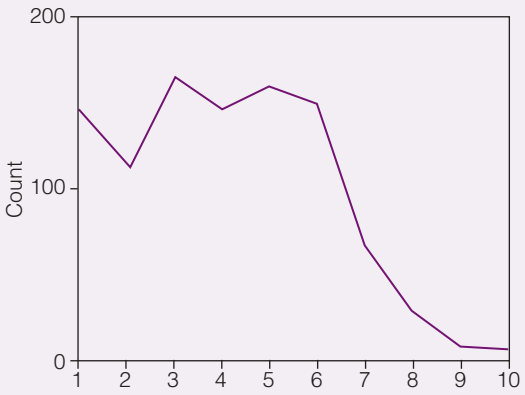


FIGURE 38.8 Staff voluntarily taking on coordination roles

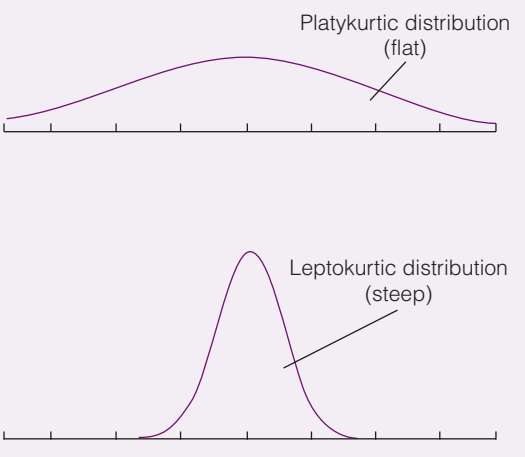


FIGURE 38.9 Types of kurtosis

it conforms to a normal curve of distribution. Another way is to conduct specific tests of skewness and kurtosis, and many statistics packages (e.g. SPSS) do this. Another way is to conduct an overall test of how normal the distribution is, for example the Shapiro-Wilk and the Kolmogorov-Smirnov tests (available in SPSS, and many researchers prefer the Shapiro-Wilk test here as being more reliable). If the distributions are too far away from a normal curve of distribution then it may be unwise to use statistics for parametric data; instead, statistics for non-parametric data should be used. On the other hand Pallant (2016) suggests that having a large sample (she suggests a size of >30) can attenuate problems of violations of normality, though other researchers adopt a more stringent position here.

The question is raised of what constitutes an acceptable level of skewness and kurtosis for a distribution to be acceptably 'normal', i.e. conforming sufficiently closely to the normal curve of distribution. There is no hard and fast rule here. Some researchers suggest that scores anywhere between –1 and +1 on skewness and on kurtosis are acceptable. Others would argue that an acceptably 'normal' degree of skewness (the figure, for example, given in software calculations of skewness) should not exceed twice the standard error of skewness, i.e. it should be within the range from minus twice the standard error of skewness to plus twice the standard error of skewness. So, for example, if the standard error of skewness (which is given in statistics software, e.g. SPSS) is 0.132 (i.e. as mild skew) then the range of acceptable skewness should be between –0.264 to +0.264; if the standard error of skewness is 0.436 then the range of acceptable skewness should be between –0.872 and +0.872. The same procedure can be adopted for calculating an acceptably 'normal' degree of kurtosis, this time working with the standard error of kurtosis (again, statistics software such as SPSS calculates this), though several researchers would not regard abnormal kurtosis as such a problem as skewness, and might even be overlooked. Small samples might be vulnerable to problems of skewness and kurtosis, and this argues for the benefits of large samples in statistics.

Normal distributions are often affected by outliers: those extreme cases/scores which are outside the normal pattern of the distribution, i.e. which are an abnormal distance from other cases/scores, often taken to be more than 1.5 times the interquartile range away from the 75th and 25th percentile. Outliers can be spotted by constructing a histogram and looking at those which do not fit the pattern, or, more technically, by constructing boxplots (see Chapter 40). Many software packages (e.g. SPSS) have a function to detect outliers and to inform the researcher of which cases/scores are outliers.

Box 38.1 provides the SPSS command sequence for calculating skewness and kurtosis.

The Shapiro-Wilk and the Kolmogorov-Smirnov tests of normality can also be used to identify outliers in SPSS (often the Shapiro-Wilk test is the preferred statistics here). The SPSS command sequence is set out in Box 38.2.

SPSS then produces output with a box marked 'Extreme Values' (Table 38.1), in which the column 'Case number' indicates those cases which are outliers, being either exceptionally/unusually low or exceptionally/unusually high. Table 38.2 presents the tests of normality. Here both the Shapiro-Wilk and the Kolmogorov-Smirnov tests indicate statistical significance (discussed in Chapter 39), and this suggests a non-normal distribution, i.e. the distributions are statistically significantly different from a normal distribution.

BOX 38.1 SPSS COMMAND SEQUENCE FOR CALCULATING SKEWNESS AND KURTOSIS

The command sequence for SPSS to calculate kurtosis and skewness is: 'Analyze' → 'Descriptive Statistics' → 'Frequencies' → Send to the 'variables' box the variables of interest → Click 'Statistics'. This opens a new window → In the 'Distributions' area, check the boxes marked 'Skewness' and 'Kurtosis' → Click 'Continue' → Click 'OK'.

BOX 38.2 SPSS COMMAND SEQUENCE FOR THE SHAPIRO-WILK AND THE KOLMOGOROV-SMIRNOV TESTS OF NORMALITY

The command sequence to calculate these is: 'Analyze' → 'Descriptive Statistics' → 'Explore' → Send to the 'Dependent List' box the variables of interest → Click 'Statistics' box → Check the 'Outliers' box → Click 'Continue' → Click the 'Plots' box → Check the boxes marked 'Normality plots with tests' and 'Histogram' → Click 'Continue' → Click 'OK'.

TABLE 38.1 EXTREME VALUES IN THE SHAPIRO-WILK TEST (SPSS OUTPUT)

Extreme Values

			Case number	*Value*
Mathematics pre-test score	Highest	1	8	10
		2	11	10
		3	159	10
		4	466	10
		5	472	10
	Lowest	1	97	2
		2	464	4
		3	208	4
		4	140	4
		5	500	5

TABLE 38.2 TESTS OF NORMALITY (SPSS OUTPUT)

Tests of Normality

	Kolmogorov-Smirnov[a]			*Shapiro-Wilk*		
	Statistic	*df*	*Sig.*	*Statistic*	*df*	*Sig.*
Mathematics pre-test score	0.222	500	0.000	0.912	500	0.000

Note
a Lilliefors Significance Correction.

38.10 Conclusion

In this chapter we have introduced some key foundations of statistical analysis:

- scales of data: nominal (categorical), ordinal, interval, ratio;
- parametric and non-parametric data, which inform the decision on whether to adopt parametric statistics or distribution-free statistics;
- descriptive and inferential statistics; inferential statistics often require conformity to the normal curve of distribution;
- kinds of variables: categorical, discrete, continuous, independent, dependent, moderator, mediator;
- hypotheses: the null hypothesis and the alternative hypothesis;
- one-tailed and two-tailed tests: those which predict the direction of results and those which do not, respectively;
- confidence intervals: the confidence that a researcher can have that his or her found result will fall within an acceptable range (interval) of results;
- distributions: the normal curve of distribution, skewness, kurtosis and the influence of outliers.

We have alerted researchers to the desirability of having as large a sample size as possible when working with statistics, as this reduces problems of standard error, confidence intervals and normal distributions. We have also alerted researchers against using parametric statistics willy-nilly, and to use them for non-parametric data or for parametric data whose distribution is insufficiently close to a normal distribution to warrant the use of those statistics which assume a normal distribution. We have indicated that many of the 'safety checks' for normality of distributions and confidence intervals are routinely available in statistics packages, and we advocate the use of such packages and such checks. Each of the bullet points above is an important consideration in deciding which statistics to use, which can be used and which should not be used. We return to this in the subsequent chapters.

Finally, the following chapters introduce some widely used Greek letters in statistics. Table 38.3 provides a brief list of widely used Greek letters.

TABLE 38.3 FREQUENTLY USED GREEK LETTERS IN STATISTICS

Greek letter	*Name*	*Use in statistics*
α	Alpha	Probability of making a Type I error. The statistical significance level
β	Beta	Probability of making a Type II error. The beta value in multiple regression is a measure of how strongly each independent (predictor) variable influences the dependent variable.
$\Delta\ \delta$	Delta	Difference: Δ. Standard deviation: δ
η	Eta	The partial regression coefficient, measure of effect size: η (usually used as 'partial eta squared' η^2)
$\Lambda\ \lambda$	Lambda	A test of mean differences in multivariate analyses (Wilks's lambda)
μ	Mu	Population mean: μ
ν	Nu	Degrees of freedom: ν
π	Pi	Population proportion: π
ρ	Rho	Correlation coefficient. Significance level: ρ
$\Sigma\ \sigma$	Sigma	The sum of: Σ. Population standard deviation (lower case: σ). Population variance: σ^2
$\Phi\ \phi\ \phi$	Phi	The phi coefficient is a measure of the degree of association between two binary variables (a binary variable has only two values, e.g. male/female).
χ	Chi	Goodness of fit and independence of two or more variables (chi-square: χ^2)

Companion Website

The companion website to the book provides additional materials, data files and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Statistical significance, effect size and statistical power

CHAPTER 39

This chapter builds on the previous chapter by introducing some key points in statistical analysis: statistical significance, effect size and statistical power. These are essential ingredients of statistics, though, as we suggest below, statistical significance has become increasingly questionable. Not only will researchers need to understand them, what they are used for and how they are used, but they will need to understand the cautions that accompany them. This chapter is designed to address these points, and it includes discussions of:

- statistical significance
- concerns about statistical significance
- hypothesis testing and null hypothesis significance testing
- effect size
- statistical power

39.1 Introduction

Though statistical significance is widely used in educational research, this chapter suggests that it has serious limitations and may not be fit for purpose in some research. It is argued that effect size can be more meaningful for educational research, and the chapter indicates the concept, practice and interpretation of effect size. Statistical significance is frequently used to test hypotheses ('null hypothesis significance testing': NHST) (Kline, 2004; Cumming, 2012), and the present chapter indicates how this operates and what are its limitations and some of its problems.

Finally, the chapter addresses an important area of statistically based research, that of statistical power. Statistical power draws on issues of sample size (see Chapter 12), statistical significance and effect size, and the discussion below shows how to put all of these together to determine how powerful a piece of research can be: how far it can find a true effect – a true positive or a true negative – and avoid a false positive or a false negative. Statistical power is an essential ingredient of quantitative research, and this chapter shows how to proceed with it.

We provide some formulae in discussing effect size; the novice researcher does not need to be put off by these as the text alone can be sufficient in introducing the issue in question.

39.2 Statistical significance

Much statistical analysis hinges on the notion of statistical significance. Kirk (1999, p. 337) indicates that 'a statistically significant result is one for which chance is an unlikely explanation'. Statistical significance purports to be a test of whether or not a result has been found by chance, a test of the 'rareness' of chance alone (Carver, 1978, p. 381).

Let us take an example from correlational research to unpack statistical significance. A correlation enables a researcher to ascertain whether, and to what extent, there is a degree of association between two variables (discussed more fully later in this chapter). Let us imagine that we observe that many people with large hands also have large feet and that people with small hands also have small feet (see Morrison, 1993, pp. 136–40). We decide to conduct an investigation to see if there is any correlation or degree of association between the size of feet and the size of hands, or whether it is just by chance that some people have large hands and large feet. We measure the hands and the feet of 100 people and observe that, 99 times out of 100, people with large feet also have large hands. Convinced that we have discovered an important relationship, we run the test on 1,000 people, and find that the relationship holds true in 999 cases out of the 1,000. That seems to be more than mere coincidence; it would seem that we could say with some certainty that if a person has large hands then she/he will also have large feet. How do we know when we can make that assertion? When do we know that we can have confidence in this prediction?

For statistical purposes, if we observe this relationship occurring 95 times out of 100, then we could say with some confidence that there seems to be a high degree of association between the two variables hands and feet, i.e. not by chance alone; no correlation would

be found in only 5 people in every 100, reported as the 0.05 level of significance (0.05 being five-hundredths). If we observe this relationship occurring 99 times out of every 100 (as in the example of hands and feet), then we could say with even greater confidence that there seems to be a very high degree of association between the two variables, i.e. not by chance alone; no correlation would be found only once in every 100, reported as the 0.01 level of significance (0.01 being one-hundredth). If we observe this relationship occurring 999 times out of every 1,000 (as in the example of hands and feet), then we could say with even greater confidence that there seems to be a very high degree of association between the two variables, i.e. not by chance alone; no correlation would be found only once in every 1,000, reported as the 0.001 level of significance (0.001 being one-thousandth).

We begin with a null hypothesis, which states that there is *no* correlation/relationship between the size of hands and the size of feet. The task is not to support the null hypothesis, i.e. the burden of responsibility is to disconfirm it. If we can show that this hypothesis is not supported for 95 per cent or 99 per cent or 99.9 per cent of the population, then we have demonstrated that there is a statistically significant relationship between the size of hands and the size of feet at the 0.05, 0.01 and 0.001 levels of significance respectively. These three levels of significance – the 0.05, 0.01 and 0.001 levels – are the levels at which statistical significance is frequently taken to have been demonstrated, usually the first two of these three levels. The researcher would say that the null hypothesis (that there is no statistically significant relationship between the two variables) has not been supported and that the level of significance observed (ρ) is at the 0.05, 0.01 or 0.001 level.

Note here that we have used the terms 'statistically significant', and not simply 'significant'; this is important, for we are using the term in a specialized way. 'Significant', as in 'statistically significant', does not mean 'important'; many inexperienced researchers incorrectly confuse these. Similarly a very high level of statistical significance does not mean that an effect (e.g. a difference, a correlation) is large, and its converse – a very low level of statistical significance – does not mean that an effect is small (Torgerson and Torgerson, 2008, p. 128; Cumming, 2012, p. 28). A high level of statistical significance (e.g. p=0.001) simply means that it is assumed that the likelihood of the found effect occurring by chance alone is very slim, and a low level of statistical significance simply means that it is assumed that the likelihood of the found effect occurring by chance alone is greater.

Let us take a second example. Let us say that we have devised a scale of 1–8 which can be used to measure the sizes of hands and feet. Using the scale we make the following calculations for eight people, and set out the results thus:

	Hand size	*Foot size*
Subject A	1	1
Subject B	2	2
Subject C	3	3
Subject D	4	4
Subject E	5	5
Subject F	6	6
Subject G	7	7
Subject H	8	8

We can observe a perfect correlation between the size of hands and the size of feet, from the person who has a size 1 hand and a size 1 foot to the person who has a size 8 hand and also a size 8 foot. There is a perfect positive correlation (as one variable increases, e.g. hand size, so the other variable – foot size – increases, and as one variable decreases, so does the other). We can use the mathematical formula for calculating the Spearman correlation (this is calculated automatically in SPSS):

$$r = 1 - \frac{6\sum d^2}{N(N^2 - 1)}$$

where d=the difference between each pair of scores, Σ=the sum of the population, and N=the size of the population. We calculate that this perfect correlation yields an index of association – a coefficient of correlation – which is +1.00.

Suppose that this time we carry out the investigation on a second group of eight people and report the following results:

	Hand size	*Foot size*
Subject A	1	8
Subject B	2	7
Subject C	3	6
Subject D	4	5
Subject E	5	4
Subject F	6	3
Subject G	7	2
Subject H	8	1

This time the person with a size 1 hand has a size 8 foot and the person with the size 8 hand has a size 1 foot. There is a perfect negative correlation (as one variable increases, e.g. hand size, the other variable – foot

size – decreases, and as one variable decreases, the other increases). Using the same mathematical formula we calculate that this perfect negative correlation yields an index of association – a coefficient of correlation – which is –1.00.

Now, clearly it is very rare to find a *perfect* positive or a perfect negative correlation; the truth of the matter is that looking for correlations will yield coefficients of correlation which lie somewhere between –1.00 and +1.00. How do we know whether the coefficients of correlation are statistically significant or not? So let us say that we take a third sample of eight people and undertake an investigation into their hand and foot size. We enter data case by case (Subject A to Subject H), indicating their hand size and then foot size. This time the relationship is less clear because the size is more mixed, for example, Subject A has a hand size of 2 and foot size of 1, Subject B has a hand size of 1 and foot size of 2 etc.:

	Hand size	*Foot size*
Subject A	2	1
Subject B	1	2
Subject C	3	3
Subject D	5	4
Subject E	4	5
Subject F	7	6
Subject G	6	7
Subject H	8	8

Using the mathematical formula for calculating the correlation statistic, we find that the coefficient of correlation for the eight people is 0.7857. Is it statistically significant or has it occurred by chance alone? From a table of significance, we read off whether the coefficient is statistically significant or not for a specific number of cases, for example:

Number of cases	*Level of significance*	
	0.05	*0.01*
6	0.93	0.96
7	0.825	0.92
8	0.78	0.875
9	0.71	0.83
10	0.65	0.795
20	0.455	0.595
30	0.36	0.47

We see that for eight cases in an investigation, the correlation coefficient has to be 0.78 or higher if it is to be statistically significant at the 0.05 level, and 0.875 or higher if it is to be statistically significant at the 0.01 level of significance. As the correlation coefficient in the example of the third experiment with eight subjects is 0.7857 we can see that it is higher than that required for significance at the 0.05 level (0.78) but not as high as that required for significance at the 0.01 level (0.875). We are safe, then, in stating that the degree of association between the hand and foot sizes does not support the null hypothesis and demonstrates statistical significance at the 0.05 level.

The first example above of hands and feet is very neat because it has 100 people in the sample. If we have more or fewer than 100 people how do we know if a relationship between two factors is statistically significant? Let us say that we have data on 30 people; in this case, because the sample size is so small, we might hesitate to say that there is a strong association between the size of hands and size of feet if we observe it occurring in 27 people (i.e. 90 per cent of the population). On the other hand, let us say that we have a sample of 1,000 people and we observe the association in 700 of them. In this case, even though only 70 per cent of the people in the sample demonstrate the association of hand and foot size, we might say that because the sample size is so large we can have greater confidence in the data than in the case of the small sample.

To ascertain statistical significance from a table the researcher can read off the significance level from a table of significance according to the sample size (or computer software calculates this automatically). There are many online sites that perform such calculations, as well as statistical packages (e.g. SPSS; GPower (a free source)), and these enable researchers to conform to conventional reporting standards, which are to report the absolute significance level (e.g. $\rho = 0.016$) as well as the relative level, for example, $\rho < 0.05$. In the selection from the table of significance for the third example above, concerning hand and foot size, the first column indicates the number of people in the sample and the other two columns indicate significance at the two levels. Hence, if we have thirty people in the sample then, for the correlation to be statistically significant at the 0.05 level, we would need a correlation coefficient of 0.36, whereas if there were only ten people in the sample, we would need a correlation coefficient of 0.65 for the correlation to be statistically significant at the same 0.05 level. Most statistical packages (e.g. SPSS) automatically calculate and report the level of statistical significance. Indeed SPSS automatically asterisks each case of statistical significance at the 0.05 and 0.01

levels or smaller. We discuss correlational analysis in more detail later in Chapter 40, and we refer the reader to that discussion.

39.3 Concerns about statistical significance

One has to be cautious in using statistical significance, as it has attracted so much criticism that some journals have ceased to accept papers that rely on null hypothesis significance testing (NHST) alone. For decades authors have gone so far as to aver that it is a discredited approach to research (Carver, 1978; Falk and Greenbaum, 1995; Ziliak and McCloskey, 2008; Gorard, 2016), even though it survives to the present. In an influential paper, Carver (1978) dismisses significance testing as a 'corrupt scientific method' (p. 387), though Cortina and Landis (2011) still see some value in it. There are several serious concerns about significance testing; we introduce key concerns here and we advise researchers to reconsider strongly the use of significance testing on its own.

The null hypothesis

Significance testing (NHST) works on the basis of commencing with the null hypothesis, seeking to support it or not to support it from the data. It claims to determine whether or not findings occur by chance. However, this relies on random sampling, but in practice such sampling occurs very rarely (Gorard, 2016). Further, even if we assume that true random sampling has occurred, this does not solve the problem of assuming the null hypothesis, as in significance testing for most of the time we simply do not know the truth of the null hypothesis, i.e. whether it is safe to make such an assumption. Indeed, in reality, it is extremely unlikely that the null hypothesis will be true; it is a straw man (Carver, 1978, p. 380; Cohen, 1994; Krueger, 2001, p. 17), a false assumption or construct. However, significance testing is built on this assumption, and this fundamentally undermines its validity. If we have no grounds for believing that the null hypothesis is true, then making such assumptions is unwise (Carver, 1978, p. 382). (Of course, there may be occasions when the null hypothesis can be demonstrated to be true, e.g. if the means and standard deviations of two groups on, say, the marks on a maths test, are shown to be the same, i.e. no difference between them.)

Statistical significance, as Carver (1978) suggests, 'simply means statistical rareness' (p. 381), not certainty. For example, accepting a significance level of 0.05 means that five times in 100 the researcher should accept the null hypothesis, and that the 'calculated risk' (p. 381) in rejecting the null hypothesis is wrong. NHST *assumes* that the null hypothesis is acceptable (Carver, 1978), but, as he remarks, this is questionable, since 'there is no way that we can be absolutely sure the null hypothesis is true. If we could be sure, we would never test for statistical significance at all' (p. 381) (see also Falk and Greenbaum, 1995; Gorard, 2016).

Though one can compute the probability of the data being statistically significant, given the assumption of the null hypothesis, the researcher cannot perform the opposite, which is to consider the likelihood of the null hypothesis, given the data (Carver, 1978; Falk and Greenbaum, 1995). In other words, logically speaking, we cannot assume the acceptability of the null hypothesis in the first place; we do not know if the initial null hypothesis is true or false (Carver, 1978; Falk and Greenbaum, 1995; Krueger, 2001; Gorard, 2016).

Ziliak and McCloskey (2008) draw attention to this error of the 'fallacy of the transposed conditional' (p. 41) in using statistical significance. In this, the 'probability of the data, given the hypothesis' is falsely transposed to be 'the probability of the hypothesis, given the data' (p. 41). Carver (1978) provides a very clear example of the danger here: the probability of obtaining a dead person, given that the person was hanged, is extremely high (e.g. 0.97 or higher), but the probability that a person has been hanged, given that he is dead, is extremely low (e.g. 0.01 or lower). It is unlikely that, in real life, the two would be confused, but in working with statistical significance, this is often exactly what happens. Put simply, it is mistaken to assume that the null hypothesis is an acceptable starting point and usually we have no way, using significance testing, of knowing if this assumption is safe; indeed it is likely to be unsafe.

If the assumption of the null hypothesis is either false or unable to be proved by the data, then this seriously questions the validity of significance testing which is based on the assumption of the null hypothesis (NHST) (Carver, 1978; Falk and Greenbaum, 1995; Gorard, 2016). Indeed Gorard (2016) show that tests of statistical significance 'are more likely to produce the wrong answer than the right one' (p. 1) and he comments that significance tests cannot even meet their supposed claim which is to indicate whether or not a finding is or is not by chance (p. 3), i.e. they are fundamentally flawed.

Statistical significance and sample size

Statistical significance is calculated as a function of sample size, and it is highly likely that statistical significance will be found if large samples are used (e.g. Ziliak and McCloskey, 2008). Statistical

significance varies according to the size of the number in the sample (see the table of significance reproduced above). In order to determine statistical significance we must have two facts in our possession: the size of the sample and, in correlational research, the coefficient of correlation or, in other kinds of research, the appropriate coefficients or data (there are many kinds, depending on the statistical test being used). Here, as the selection from the table of significance reproduced above shows, the coefficient of correlation can decrease and still be statistically significant as long as the sample size increases. (This resonates with Krejcie's and Morgan's (1970) principles for sampling, observed in Chapter 12, namely, as the population increases, the sample size increases at a diminishing proportion in addressing randomness.)

This is a major source of debate for critics of statistical significance and NHST, who argue that it is almost impossible *not* to find statistical significance when dealing with large samples, as the coefficients can be very low and still attain statistical significance. Statistical significance might be easy to find with large samples, even though the size of the difference or the correlation might be very small indeed (Cumming, 2012, p. 29). Statistical significance varies with sample size and brings with it the possibility of a Type II error (a false negative) if only significance testing is used, particularly with small samples (e.g. Torgerson and Torgerson, 2008; Ziliak and McLoskey, 2008; Ellis, 2010).

Statistical significance on its own has come to be seen as an unacceptable index of effect (Thompson and Snyder, 1997; Wilkinson and the Task Force on Statistical Inference, APA Board of Scientific Affairs, 1999; Thompson, 2002; Wright, 2003; Kline, 2004; Ziliak and McLoskey, 2008; American Psychological Association (APA), 2010; Ellis, 2010), for the reasons given above and because it depends on both sample size and the coefficient (e.g. of correlation). Statistical significance can be attained *either* by having a large coefficient together with a small sample *or* having a small coefficient together with a large sample. The problem is that one is not able to deduce which is the determining effect from a study using statistical significance (Coe, 2000, p. 9). It is important to be able to tell whether it is the sample size or the coefficient that is making the difference.

False dichotomies

Statistical significance is seen as arbitrary in its cut-off points and an unhelpful obstacle rather than a facilitator in educational research (even though the 0.05 significance level corresponds to approximately two standard deviations above the mean and the 0.01 significance level corresponds to approximately three standard deviations above the mean). One also has to be cautious in null hypothesis significance testing (NHST), as it may encourage dichotomous thinking, i.e. a finding is or is not statistically significant, because of these arbitrary cut-off points, and this may discourage thinking of alternative ways of testing a hypothesis (Kline, 2004, pp. 76–9).

Statistical and educational significance

Statistical significance is not the same as educational significance (cf. Ziliak and McCloskey, 2008, p. 110). For example, I might find a statistically significant correlation between the amount of time spent on mathematics and the amount of time spent shopping. This may be completely unimportant. Significance testing does not say anything about whether a finding is highly likely, or important or trivial (Carver, 1978).

Similarly I might find that there is no statistically significant difference between males and females in their liking of physics. However, closer inspection might reveal that there is a difference. Say, for example, that males prefer physics to females, but that the difference does not reach the 'cut-off' point of the 0.05 level of significance; maybe it is 0.065. To say that there is no difference or simply to support the null hypothesis here might be inadvisable. There are two issues here: (a) the cut-off level of significance is comparatively arbitrary, though high; (b) one should not ignore coefficients that fall below the conventional cut-off points. This leads us into a discussion of effect size as an alternative to significance levels (discussed below). Statistical significance, as Ziliak and McCloskey (2008) and Ellis (2010) note, is a putative indicator of chance, for example, that something exists by chance and not by chance (but in light of all the cautions that we set out above), whilst effect size is an indicator which has greater practical significance as researchers may be less interested in chance or proof of existence and more interested in size. We discuss effect size below.

Utility value for research

Statistical significance says nothing about what many researchers really want to know: the size of an effect (e.g. the *amount* of difference or correlation) (Ziliak and McLoskey, 2008; Ellis, 2010); how much. A measure of effect size may be more useful than statistical significance. Statistical significance on its own is no indication of impact, and impact is what researchers (and politicians and funding bodies) are keen to establish.

Given these concerns about statistical significance, we suggest that researchers note where significance testing is and is not fit for purpose and address possible alternatives to significance testing. These include, for example, effect size, statistical power (discussed below)

and counterfactual analysis, i.e. the number of 'counterfactual cases needed in order to disturb a finding' (Gorard and Gorard, 2015, p. 484).

39.4 Hypothesis testing and null hypothesis significance testing

The example that we gave above, of correlational analysis, illustrates a wider issue of NHST (Kline, 2004; Ellis, 2010; Cummings 2012), so we introduce here how this might proceed and then we inject a note of great caution into the use of NHST. NHST has been widely used, and it is for this reason that we retain it in this chapter, but we recognize the sometimes serious problems that inhere in it, and we argue that hypothesis testing can, indeed should, proceed without reliance on NHST. NHST can follow four stages, set out below. However, we strongly counsel readers to question the assumption that hypothesis testing relies on statistical significance testing alone: the two are not the same and, as we have indicated above, NHST has problems and limitations that are so serious as to discredit it in the eyes of many researchers.

Stage 1

In quantitative research, as mentioned above, we commence with a null hypothesis, for example:

- there is *no* statistical significance in the distribution of the data in a contingency table (crosstabulation);
- there is *no* statistically significant correlation between two factors;
- there is *no* statistically significant difference between the means of two groups;
- there is *no* statistically significant difference between the means of a group in a pre-test and a post-test;
- there is *no* statistically significant difference between the means of three or more groups;
- there is *no* statistically significant difference between two sub-samples;
- there is *no* statistically significant difference between three or more sub-samples;
- there is *no* significant prediction capability between one independent variable X and dependent variable Y;
- there is *no* significant prediction capability between two or more independent variables X, Y, Z … and dependent variable A.

The task of the research is to support or not to support the null hypothesis. We remind readers here that the starting point of NHST, which assumes the null hypothesis to be true, in fact is questionable or even false (Kline, 2004, p. 70; Ziliak and McCloskey, 2008).

Stage 2

Having set the null hypothesis, the researcher then sets the level of significance (α) that will be used to support or not to support the null hypothesis; this is the alpha (α) level. The level of alpha is determined by the researcher. Typically it is 0.05, i.e. the chance of a finding being by chance alone – the null hypothesis being supported – is 5 per cent. In writing this we could say 'Let $\alpha=0.05$'. If one wished to be more robust then one would set a higher alpha level ($\alpha=0.01$ or $\alpha=0.001$). This is the level of risk that one wishes to take in supporting or not supporting the null hypothesis.

Stage 3

Having set the null hypothesis and the level at which it will be supported or not supported, one then computes the data as appropriate for the research in question (e.g. measures of association, measures of difference, regression and prediction measures).

Stage 4

Having analysed the data one is then in a position to support or not to support the null hypothesis, and this is what is reported.

In hypothesis testing one has to avoid Type I and Type II errors (see also Chapter 12). A Type I error occurs when one does not support the null hypothesis when it is in fact true (a false positive). This is a particular problem as the sample size increases, as the chances of finding a statistically significant association increases; to overcome this one can set a higher alpha (α) limit (e.g. 0.01 or 0.001) for statistical significance to be achieved. A Type II error occurs when one supports the null hypothesis when it is in fact not true (a false negative), which is often the case if the levels of significance are set too stringently; to overcome this the researcher can set a lower alpha level (α) (e.g. 0.1 or 0.2). Type I and Type II errors are represented in Table 39.1.

In considering hypothesis testing, however, given the serious questions which we have raised earlier

TABLE 39.1 TYPE I AND TYPE II ERRORS

Decision	*H_0 true*	*H_0 false*
Support H_0	Correct	False negative Type II error (β)
Do not support H_0	False positive Type I error (α)	Correct

against NHST, we suggest here that hypothesis testing moves away from NHST and towards alternative ways of hypothesis testing, for example, effect size, statistical power, falsification (in the tradition of Popper) and counterfactual analysis, conducting replication studies, using mixed methods and triangulation, ruling out alternative hypotheses and theories, and giving weight to the power of evidence, rather than simply to a questionable reliance on significance testing. Hypothesis testing is not the same as null hypothesis significance testing.

39.5 Effect size

Statistical significance only purports to tell the researcher whether a particular result (e.g. a difference, a correlation) has or has not occurred by chance. That is all, and even then, as we saw above, the assumptions of the null hypothesis and of NHST are highly questionable. How many researchers are actually interested in whether something does or does not occur by chance, even if the assumptions underpinning NHST are true? Most researchers are more concerned with the size, the magnitude, of an effect, be it, for example, a difference or an association. As Ziliak and McLoskey (2008) note, statistical significance is not only a 'sizeless stare' but omits the very thing in which researchers are interested: how much.

What is required either to accompany or replace statistical significance is information about *effect size* (Wilkinson and the Task Force on Statistical Inference, APA Board of Scientific Affairs, 1999; Kline, 2004; APA, 2010). Indeed effect size is so much more important than statistical significance that, as indicated earlier, many international journals have either abandoned statistical significance in favour of reporting effect size, or have insisted that statistical significance be accompanied by indications of effect size (Thompson, 2002; APA, 2010). Some measures of effect size are standardized, for example, Cohen's *d*, regression weights (β); others use the original units, for example, means, medians, regression weights (*b*); and others are unit-free, for example, percentages, correlation coefficients, proportion of explained variance.

Effect size is a measure of magnitude: the amount of something; how much (e.g. Cohen 1988; Cumming, 2012, p. 34). It operates in two spheres: (i) measures of difference and (ii) measures of association (cf. Kline, 2004, p. 97; Ellis, 2010, p. 7; Cumming, 2012). It is in fact an inaccurate term as 'effect' implies causality, whereas none is actually inferred. With regard to difference, effect size is a way of quantifying the difference between two or more groups. For example, with two groups, if an experimental group has received an intervention whilst the control group has not, then the effect size is a measure of *how big* the effect/difference is between the two groups, which is something that statistical significance does not tell us (Coe, 2000; Wright, 2003, p. 125).

There are many ways of calculating effect size. Glass *et al.* (1981) calculates the effect size as:

$$\frac{\text{(mean of experimental group – mean of control group)}}{\text{standard deviation of the control group}}$$

Coe (2000, p. 7), whilst acknowledging that there is a debate on whether to use the standard deviation (SD) of the experimental or control group as the denominator, suggests that the SD of the control group is preferable as it provides 'the best estimate of standard deviation, since it consists of a representative group of the population who have not been affected by the experimental intervention'. Many calculations of effect size use a 'pooled' estimate of standard deviation, as this is more accurate than that provided by the control group alone. To calculate the pooled deviation Coe provides the formula:

$$SD\ pooled = \sqrt{\frac{(N_E - 1)SD_E^2 + (N_C - 1)SD_C^2}{N_E + N_C - 2}}$$

where N_E = number in the experimental group, N_C = number in the control group, SD_E = standard deviation of the experimental group and SD_C = standard deviation of the control group.

The formula for the pooled deviation is (Ellis, 2010; Cumming, 2012):

$$\frac{\text{(mean of experimental group – mean of control group)}}{\text{pooled standard deviation}}$$

where the pooled standard deviation = (standard deviation of group 1 + standard deviation of group 2).

Morris (2008) suggests that effect size can be calculated thus:

$$\frac{(\mu_{T,post}(\mu_{T,pre}))(\mu_{C,post}(\mu_{C,pre}))}{\text{pooled standard deviation}}$$

where:

$\mu_{T,post}$ = the post-test mean of the treatment (experimental) group

$\mu_{T,pre}$ = the pre-test mean of the treatment (experimental) group

$\mu_{C,post}$ = the post-test mean of the control group

$\mu_{C,pre}$ = the pre-test mean of the control group

This is attractive, as it takes account of changes over time in both groups (see also Figure 20.1 and the 'subtraction' method set out in Chapter 20).

There are several different calculations of effect size, and we list these in Table 39.2, together with summary guidelines on whether these are small, medium or large. However, these guidelines are not absolute; they are *guideline*s only, and are not fixed and/or immutable. Nevertheless Cohen (1988) indicated that these guidelines are useful for researchers. Some researchers (e.g. Kline, 2004; Ellis, 2010; Cumming, 2012) argue against simply 'reading off' effect sizes into 'small', 'medium' and 'large', and they suggest that context and comparative value also have to be taken into account.

For difference tests, Cohen's *d* is the most widely used. Glass's delta can be used if group sizes in the sample are different or if the standard deviation of the groups is different. Hedges' *g* can be used if the group sizes in the sample are unequal or if the sample size is small. These tests report in standardized units, i.e. 0.5 means a difference of half of a standard deviation. This is useful in comparing studies which may have used different scales or units.

Different kinds of statistical treatments use different effect size calculations. Many calculations of effect size give an estimate between 0 and 1; other formulae can yield an effect size that is larger than 1 (see Coe, 2000). In using Cohen's *d*:

0–0.20 = weak effect
0.21–0.50 = modest effect
0.51–1.00 = moderate effect
>1.00 = strong effect

In correlational data the coefficient of correlation is used as the effect size in conjunction with details of the direction of the association (i.e. a positive or negative correlation). The coefficient of correlation (effect size) is interpreted thus:

<0 +/–1 weak
<0 +/–3 modest
<0 +/–5 moderate
<0 +/–8 strong
≥ +/–0.8 very strong

We provide more detail on interpreting correlation coefficients later in this chapter. Thompson (2001, 2002) argues forcibly against simplistic interpretations of effect size as 'small', 'medium' and 'large', as to do this commits the same folly of fixed benchmarks as that of statistical significance, i.e. 'we would merely be being stupid in another metric' (Thompson, 2001, pp. 82–3). Rather, he avers, it is important to avoid fixed benchmarks (i.e. cut-off points), and relate the effect sizes found to those of prior studies, confidence intervals and power analyses (discussed below). We discussed the *confidence interval* in Chapters 12 and 38; it is reported as, for example 90 per cent, 95 per cent, 99 per cent, and is calculated as 1–α, i.e. the level of confidence that one can have in the view that a score falls within a prespecified range of scores (e.g. 95 per cent, 99 per cent) (Ellis, 2010). Software for calculating confidence intervals for many measures can be found at:

- www.surveysystem.com/sscalc.htm
- www.danielsoper.com/statcalc3/calc.aspx?id=96.

TABLE 39.2 EFFECT SIZES FOR DIFFERENCE AND ASSOCIATION

Type	*Statistic*	*Effect sizes*		
		Small	*Medium*	*Large*
Difference testing (t-tests)	Cohen's d (*d*)	0.20	0.50	0.80
	Glass's delta (*d*)	0.20	0.50	0.80
	Hedges' (*g*)	0.20	0.50	0.80
Correlation analysis	Pearson's *r* (scale data)	0.10	0.30	0.50
	Spearman's rho (r_s) (ordinal data)	0.10	0.30	0.50
Crosstabulation (correlation)	Phi (φ) coefficient (2×2 crosstabs)	0.10	0.30	0.50
	Cramer's V (any size crosstabs)	0.10	0.30	0.50
Multiple regression (correlation)	R^2	0.02	0.13	0.26
	Adjusted R^2 ($^{adj}R^2$)	0.02	0.13	0.26
ANOVA (correlation)	Eta squared (η^2)	0.01	0.06	0.14
	Cohen's *f*	0.10	0.25	0.40
MANOVA (correlations)	Partial eta squared (η^2)	0.01	0.06	0.14

Wright (2003, p. 125) also suggests that it is important to report the units of measurement of the effect size, for example in the units of measurement of the original variables as well in standardized units (e.g. standard deviations), the latter being useful if different scales of measures are being used for the different variables.

In calculating the effect size (eta squared) for independent samples in a t-test (see Chapter 41), the following formula can be used:

$$\text{Eta squared} = \frac{t^2}{t^2 + (N_1 + N_2 - 2)}$$

Here t = the t-value (calculated by SPSS); N_1 = the number in the sample of group one and N_2 = the number in the sample of group 2.

Let us take an example of the results of an evaluation item to see how large is the difference in the voting between two groups – (a) leaders/senior managers (SMT) of schools, and (b) teachers – on the item 'How well learners are cared for, guided and supported', referring to Tables 39.3 and 39.4.

Here the t-value is 1.923, N_1 is 347 and N_2 is 653. Hence the formula is:

$$\text{Eta squared} = \frac{t^2}{t^2 + (N_1 + N_2 - 2)} = \frac{1.923^2}{1.923^2 + (347 + 653 - 2)} = \frac{3.698}{3.698 + 998} = 0.0037$$

Using the guidance on interpreting effects sizes above, the result of 0.003 is a tiny effect size, i.e. only 0.3 per cent difference between the two groups on the item 'How well learners are cared for, guided and supported'.

For a paired sample t-test (Chapter 41) the effect size (eta squared) is calculated by the following formula:

$$\text{Eta squared} = \frac{t^2}{t^2 + (N_1 - 1)}$$

As another example, let us imagine that the same group of students had scored marks out of 100 in 'Maths' and 'Science' (Tables 39.5 and 39.6).

The effect size can be worked out thus (using SPSS):

$$\text{Eta squared} = \frac{t^2}{t^2 + (N_1 - 1)} = \frac{16.588^2}{16.588^2 + (1000 - 1)} = \frac{275.162}{275.162 + 999} = 0.216$$

In this example the effect size is 0.216, a large effect, i.e. there was a substantial difference between the scores of the two groups.

For Analysis of Variance (discussed in Chapter 41) the effect size is calculated thus:

$$\text{Eta squared} = \frac{\text{Sum of squares between groups}}{\text{Total sum of squares}}$$

In SPSS this is given as 'partial eta squared'. For example, let us imagine that we wish to compute the effect size of the difference between four groups of schools on mathematics performance in a public examination. The four groups of schools are: (a) rural primary; (b) rural secondary; (c) urban primary; (d) urban secondary. Analysis of Variance yields the following result (Table 39.7):

Working through the formula yields the following:

$$\text{Eta squared} = \frac{\text{Sum of squares between groups}}{\text{Total sum of squares}} = \frac{7078.619}{344344.8} = 0.021$$

The figure of 0.021 indicates a small effect size, i.e. that there is a small difference between the four groups

TABLE 39.3 MEAN AND STANDARD DEVIATION IN AN EFFECT SIZE (SPSS OUTPUT)

Group Statistics

	who are you	N	Mean	Std. Deviation	Std. Error Mean
How well learners are cared for, guided and supported	leader/member of the SMT	347	8.37	2.085	.112
	teachers	653	8.07	2.462	.096

TABLE 39.4 THE LEVENE TEST FOR EQUALITY OF VARIANCES (SPSS OUTPUT)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
How well learners are cared for, guided and supported	Equal variances assumed	8.344	.004	1.923	998	.055	.30	.155	-.006	.603
	Equal variances not assumed			2.022	811.922	.044	.30	.148	.009	.589

TABLE 39.5 MEAN AND STANDARD DEVIATION IN A PAIRED SAMPLE TEST (SPSS OUTPUT)

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	MATHS	81.71	1000	23.412	.740
	SCIENCE	67.26	1000	27.369	.865

TABLE 39.6 DIFFERENCE TEST FOR A PAIRED SAMPLE (SPSS OUTPUT)

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	MATHS - SCIENCE	14.45	27.547	.871	12.74	16.16	16.588	999	.000

TABLE 39.7 EFFECT SIZE IN ANALYSIS OF VARIANCE (SPSS OUTPUT)

ANOVA

MATHS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7078.619	3	2359.540	4.205	.006
Within Groups	337266.2	601	561.175		
Total	344344.8	604			

in their mathematics performance (note that this is a much smaller difference than that indicated by the significance level of 0.006, which suggests a highly statistically significant difference between the four groups of schools).

In regression analysis (discussed in Chapter 42) the effect size of the predictor variables is given by the beta weightings. In interpreting effect size here Muijs (2004, p. 194) gives the following guidance:

0–0.1	weak effect
0.1–0.3	modest effect
0.3–0.5	moderate effect
>0.5	strong effect

For a discussion of the importance of attending to both small and large effect sizes, see Wang (2008). Wang argues that small effect sizes could indicate an important finding (p. 130), i.e. effect size and importance are two separate concepts.

Hedges (1981) and Hunter *et al.* (1982) suggest alternative equations to take account of differential weightings due to sample size variations. The two most frequently used indices of effect sizes are standardized mean differences and correlations, though with non-parametric statistics, for example, the median, can be used. Lipsey (1992, pp. 93–100) sets out a series of statistical tests for working on effect sizes, effect size means and homogeneity.

Muijs (2004, p. 126) indicates that a measure of effect size for crosstabulations, instead of chi-square, should be *phi*, which is the square root of the calculated value of chi-square divided by the overall valid sample size. For example, if chi-square is 23.716 and the sample size is 900, then *phi*=23.716/900=0.02635, and then take the square root of this=0.1623.

Effect sizes are susceptible to a range of influences. These include (Coe, 2000):

- *restricted range*: the smaller the range of scores, the greater the possibility of a higher effect size, therefore it is important to use the pooled standard deviation (not just that of one group) in calculating the effect size. It is important to report the possible restricted range or sampling here (e.g. a group of highly able students rather than, for example, the whole ability range);
- *non-normal distributions*: effect size usually assumes a normal distribution, so any non-normal distributions should be reported;
- *measurement reliability*: the reliability (accuracy, stability and robustness) of the instrument being used (e.g. the longer the test, or the more items that are used to measure a factor, the more reliable it could be).

The researcher also has to keep in mind that effect size operates at the group level rather than the individual level, and that, thereby, individual differences may be overlooked. Researchers also have to be mindful that small samples may not be able to detect a small effect size, thereby committing a Type II error (a false negative) (Torgerson and Torgerson, 2008, p. 128); the advice here, then, is to have as large a sample as possible. Additionally the researcher can set the beta level (β) at a lower significance level in order to avoid a Type II error.

There are downloadable software programs available that calculate effect size straightforwardly, and we indicate these in the companion website. More information on effect sizes can be found in Leech and Onwuegbuzie (2004), Kline (2004) and Ellis (2010).

39.6 Statistical power

For any test to be worth its salt, it is important for it to have strong statistical power. Statistical power is the probability that a study will detect an effect where it exists and will not find an effect when none exists, i.e. find a true positive and a true negative and avoid a false positive and false negative. A large sample helps the researcher to achieve statistical power.

Addressing statistical power takes us into consideration of Type I and Type II errors. A Type I error is a false positive: rejecting the null hypothesis (H_0) when, in reality, it is true. A Type II error is a false negative, accepting the null hypothesis (H_0) when, in reality, it is false. Decisions to accept or reject a null hypothesis are usually made on the basis of statistical significance (discussed earlier).

The probability of committing a Type I error (false positive) is termed 'alpha' (α) (the significance level of a test). Alpha can range from 0 to 1, and should be <0.05 to avoid a Type I error (i.e. only a 5 per cent chance of making the error). A low alpha (e.g. ≤0.05) indicates statistical significance in conventional terms, and the closer it is to 0, the lower the chance of a Type I error. A high alpha (e.g. 0.65) suggests no real statistical significance (i.e. the result is by chance).

The probability of committing a Type II error (false negative) is termed 'beta' (β). If the statistical power is high then the probability of making a Type II error (i.e. concluding that there is no effect when, in fact, there is an effect) goes down. Beta can range from 0 to 1 and should be >0.05 to avoid a Type II error (i.e. only a 5 per cent chance of making the error). A low beta (e.g. ≤0.05) indicates statistical significance, and the closer it is to 1, the lower the chance of a Type II error.

If we decrease alpha then beta will increase, and if we increase alpha then beta will decrease. If we choose

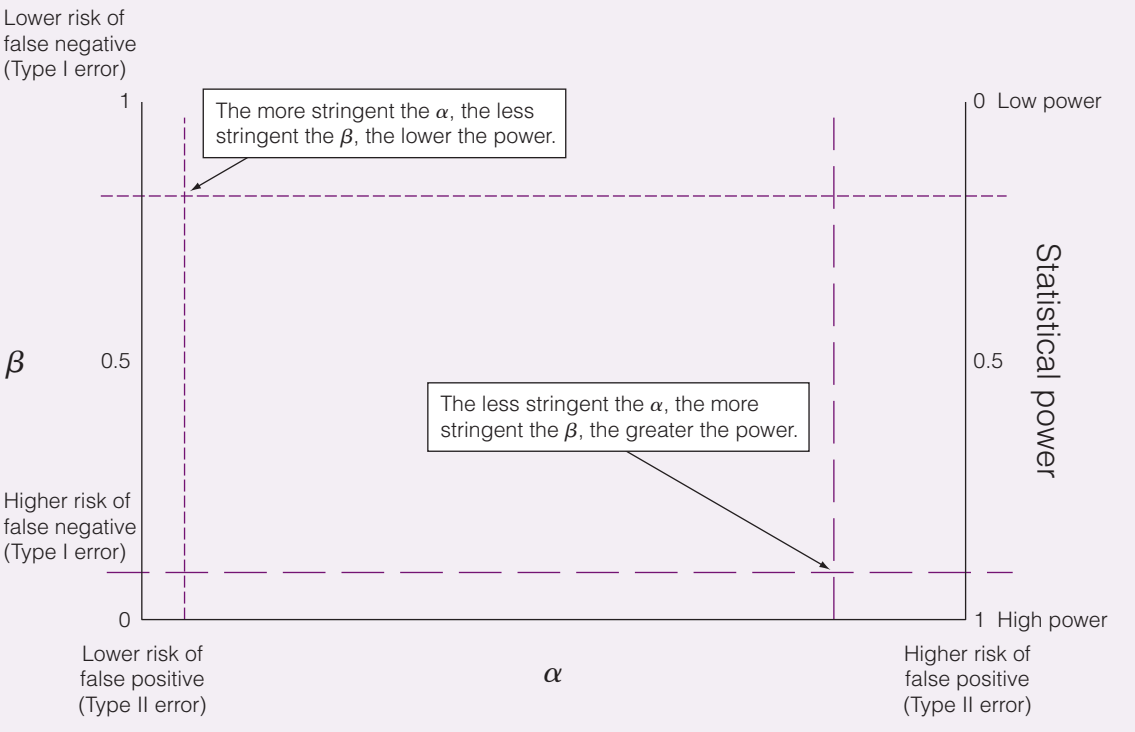
a very small alpha (e.g. $\alpha=0.001$), then we make it difficult to reject the null hypothesis, i.e. we may be making a Type II error: failing to find an effect which is actually present (a false negative). If we choose a large alpha (e.g. $\alpha=0.25$), i.e. easier to reject the null hypothesis, then we reduce the chance of making a Type II error but increase the chance of making a Type I error (false positive). If we choose a very small beta (e.g. $\beta=0.05$) then we may commit a Type I error: finding an effect that is not really present (false positive). If we choose a large beta (e.g. $\beta=0.30$), then we reduce the chance of a Type II error.

Statistical power ranges from 0 to 1, and the closer it is to 1, the greater the statistical power, and the greater the power, the higher the decimal fraction, i.e. 0.80 is more powerful than 0.50. Ziliak and McCloskey (2008) remark that high power is good whilst low power is bad (p. 132). A power of 0.50 (a 50 per cent chance of making a Type II error) has less statistical power than a power of 0.80. Beta is calculated as $\beta=1$ – power, and the power of a test is $1-\beta$. In other words, power is inversely related to beta (the probability of making a Type II error). A power level of 0.50 or lower is problematic, as a power level of 0.50 means that there is a 50/50 chance of rejecting a true null hypothesis (a false negative) or accepting a false alternative hypothesis (a false positive) (Kline, 2004, p. 43).

The relationship between α, β and statistical power is set out in Figure 39.1. Here one can see a trade-off: the more one wishes to avoid a Type I error, the greater is the chance of committing a Type II error, and the more one wishes to avoid a Type II error, the greater is the chance of committing a Type I error.

The researcher will have to make several decisions here, for example:

- what to set as the appropriate level of the alpha (α) in order to avoid a Type I error (the lower the alpha, the more rigorous and stringent is the test);
- what to set as the appropriate level of the beta (β) in order to avoid a Type II error (the lower the beta, the greater the chance of finding a true positive);
- whether to set a stringent α (e.g. 0.01), which may have low power but decreases the chance of a Type I error;



Figures for α and β are significance levels

FIGURE 39.1 Balancing alpha, beta and statistical power

- whether to set a less stringent α (e.g. 0.10), which may have high power but increases the chance of committing a Type II error;
- whether to set a stringent β (e.g. 0.05), which may have high power but increases the chance of committing a Type I error;
- whether to set a less stringent β (e.g. 0.8), which may have low power but decreases the chance of committing a Type II error.

The power level is, in part, a function of the alpha. If the researcher chooses a very small value of α (e.g. 0.05), this makes it very difficult to reject the null hypothesis but much easier to commit a Type II error (a false negative). If the researcher chooses a larger value of α (e.g. 0.10), this makes it easier to reject the null hypothesis but more difficult to commit a Type II error.

Researchers have to set the power level for themselves, deciding the power level required, and, *inter alia*, this affects the sample size. Power analyses are usually run before a study is conducted (often to determine the sample size needed). How can the researcher proceed here? The researcher looks at the relationship between α, β and power. One can set the α very stringently (e.g. 0.05) but set the β less stringently (e.g. 0.20). Cohen (1988) held that Type I errors should be treated four times more seriously than Type II errors: a four to one weighting of beta to alpha. Researchers often set an α of 0.05, and a β of 0.20, giving a power of 0.80 (80 per cent power level), i.e. a 5 per cent chance of a Type I error and a 20 per cent chance of a Type II error, thereby setting a good likelihood of finding a true positive. This is shown in Figure 39.2.

Statistical power analysis has four main parameters:

1. the effect size;
2. the sample size (number of observations);
3. the alpha (α) significance level (usually 0.05 or lower);
4. the power of the statistical test (setting the acceptable β level and the desired power (1 – β), e.g. β of 0.20 and power of 0.80).

Larger effect sizes are easier to detect than smaller effect sizes, and in larger samples effects are easier to detect than in smaller samples (Ellis, 2010). Statistical power, then, influences sample size. Further, it can be affected by the variation in the population: the greater the heterogeneity (variation), the lower the power, and the calculation of statistical power depends on the test used and on whether it is one-tailed or two-tailed (see also Chapter 12). Cohen (1988) and Ellis (2010) provide tables for determining sample sizes depending on the power required, and indeed there are many web sites that provide a similar service.

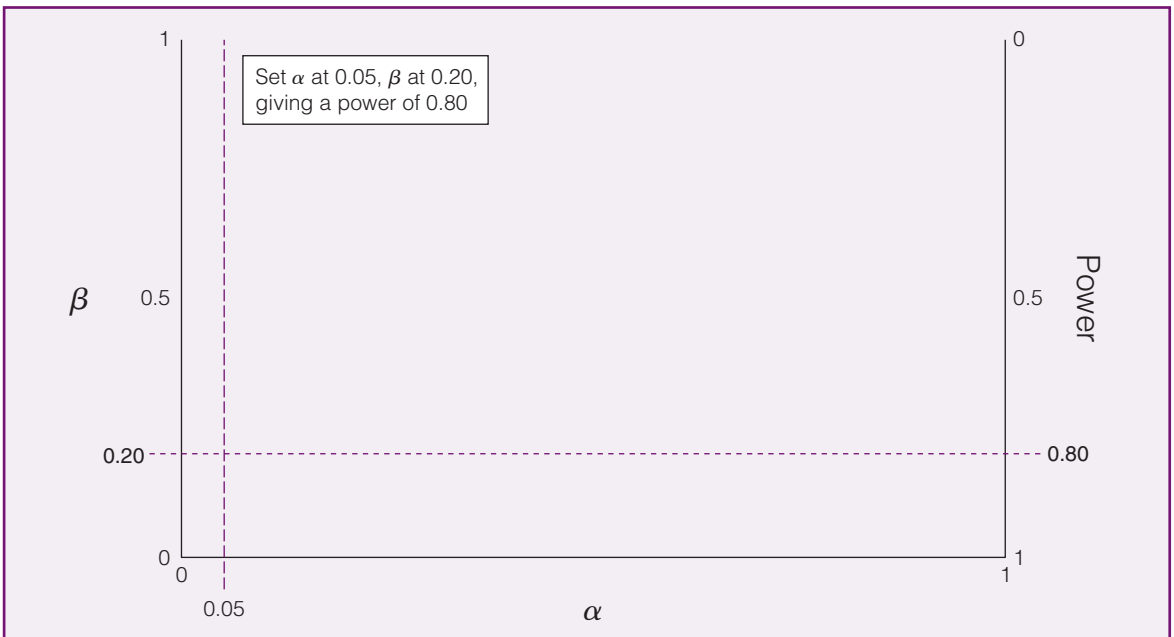


FIGURE 39.2 Setting the alpha, beta and power size

Overall, to improve the statistical power of the test, researchers should strive for the following (Kline, 2004; Ellis, 2010; Cumming, 2012):

- use a large sample;
- look for a larger effect size;
- lower the α level, as this increases the chance of rejecting the null hypothesis, i.e. reducing the chance of a Type II error;
- use a homogeneous sample;
- use a one-tailed test and ensure that the direction of the alternative hypothesis is the same as the direction of the population effect;
- ensure high reliability scores; and
- use parametric tests rather non-parametric tests (where appropriate).

Tables of sample size for statistical power, with an α of 0.05, and a power level of 0.80 suggest the following:

- For a one-tailed difference test, seeking an effect size of 0.5, the sample size is approximately 100.
- For a two-tailed difference test, seeking an effect size of 0.5, the sample size is approximately 130.
- For a one-tailed correlation, seeking an effect size of 0.5, the sample size is approximately 20.
- For a two-tailed correlation test, seeking an effect size of 0.5, the sample size is approximately 30.

These are gross approximations (i.e. heavily rounded), and only by way of indication, so readers are very strongly advised to go to tables of more exact sample sizes by power levels, alpha levels and effect size sought (see Cohen, 1988; and Ellis, 2010).

39.7 Conclusion

Statistical significance, effect size and statistical power carry great weight in educational research. However this chapter has deliberately cast a serious doubt over significance testing other than to demonstrate that a result is or is not by chance, and it has suggested that even this is highly questionable. It has argued that null hypothesis significance testing (NHST), though widely used, should be regarded with caution not only because it has serious shortcomings, for example, in its questionable assumption of the null hypothesis, but also because it does not address issues in which researchers are typically interested – the magnitude of an effect, be that effect a matter of difference or association (the two main types of effect size). Hence we have argued for the importance of calculating effect size in quantitative educational research. Measures of effect size, be they in terms of standardized units, original units or unit-free measures, vary according to the statistical tests used to calculate them. We have also indicated the possibility of using 'the number of counterfactual cases need to disturb a finding' (Gorard and Gorard, 2015, p. 484) as an alternative to significance testing. We have suggested that hypothesis testing does not commit the researcher to using only NHST, as there are other ways of testing a hypothesis.

However this is not to jettison significance testing altogether. We have indicated how it is used in determining statistical power, and that statistical power – the probability that the research will detect a true effect (a true positive) where there is one and will not find an effect where none exists (a true negative) – is a key factor in judging how important the research might be and how reliable is the instrument used. Statistical power is an essential, if often overlooked, ingredient of high-quality educational research. We have indicated that it should be decided in advance of the research and that it draws on effect size, sample size and statistical significance (in terms of alpha (α) and beta (β)), and that a trade-off has to be made between alpha and beta in addressing Type I and Type II errors, requiring the researcher to set appropriate alpha and beta levels. We have suggested that many researchers set statistical power at 0.80, alpha at 0.05 and beta at 0.20. Researchers should consult tables of sample size that take into account alphas, betas, effect size sought and statistical power.

These three elements of quantitative research – significance testing, effect size and statistical power analysis – combine with the issues raised in Chapter 38 to ensure rigorous and reliable quantitative research.

Companion Website

The companion website to the book provides additional material, data files and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Descriptive statistics

CHAPTER 40

This chapter introduces descriptive statistics. Descriptive statistics do what they say: they describe, so that researchers can then analyse and interpret what these descriptions mean. This chapter introduces some key descriptive statistics and how to use them. This includes:

- a cautionary note about missing data
- frequencies, percentages and crosstabulations
- measures of central tendency and dispersal
- taking stock
- correlations and measures of association
- partial correlations
- reliability

Descriptive statistics include frequencies, measures of dispersal (standard deviation), measures of central tendency (means, modes, medians), standard deviations, crosstabulations and standardized scores. With the exception of standardized scores, which we keep for Chapter 42, we address all of these in this chapter. In this chapter we refer to the Statistical Package for the Social Sciences (SPSS) in many of the calculations and examples.

40.1 Missing data

For all the statistics introduced in this and subsequent chapters, we alert the reader to the issue of missing data. Missing data can damage the precision or correctness of analysis and some statistics assume that there are no missing data.

There are many reasons why data might be missing. For example, respondents might simply have overlooked entering a result, or they might have deliberately left a survey answer blank, or they may drop out of the research (attrition) and so on. For whatever reason, data frequently are missing. Data may be Missing Completely At Random (MCAR) (Rubin, 1976), i.e. there is no pattern to the missing data for any variables. Data may be Missing At Random (MAR) (*ibid.*), where there is a pattern to the missing data, but not for the main dependent variable. Data may be Missing Not At Random (MNAR) (*ibid.*), where there is a pattern in the missing data that affects the main dependent variable (e.g. low-income families may not respond to a survey item).

What is the researcher to do here? It is impossible and maybe very dangerous to guess what data the respondent would have entered. Of course, the researcher could try to go back to the respondent and ask him/her to complete the missing item, but often this is impossible.

The researcher will need to ascertain how many cases for each variable have missing data (SPSS automatically calculates this using its 'Descriptives' function) and what the distribution of the missing values is, for example, whether the missing data are randomly scattered or whether there is a systematic pattern in the missing data (again, SPSS can indicate this). If the missing data are randomly scattered, then, provided that the number of missing cases is so small that it is impossible for the results to seriously distort the overall findings, the researcher might simply exclude those cases (SPSS can do this automatically). If the missing data are not randomly scattered but are systematically missing, i.e. if there is a pattern in the non-response, then this presents a major problem for the researcher, as it is impossible to determine what result/data the respondents would have given if they had given a response. In this case the researcher may decide not to pursue that part of the analysis or may use imputation methods (see below).

To decide whether the number of missing cases in a variable is sufficiently high to seriously distort the results, the researcher can conduct a sensitivity analysis (Gorard, 2013, p. 88). This involves calculating the number of *different* responses/cases (i.e. different from the non-missing data) that would be required to overturn or seriously change the findings of the analysis. If the number is so low that it could not upset the findings then the researcher might wish to proceed, reporting the number of missing cases.

The researcher can adopt a *deletion method* for missing data, excluding any cases (e.g. people) whose data are incomplete on *any* variable (SPSS does this in

its 'Exclude cases listwise' function, only using those cases which are complete on *all* the variables). Alternatively, the researcher can exclude those cases which are incomplete on only the variables of interest for a specific statistical calculation (SPSS does this in its 'Exclude cases pairwise' function). For example, if we want to calculate the difference between the scores on two tests which students have taken – say mathematics and English – then, if some students have completed the mathematics test but not the English test, or *vice versa*, then those students are excluded from the calculation and this is reported in the research.

The cost of the deletion method is in the power of the analysis, as the number of cases is reduced. For the exclusion of every incomplete case (person) (i.e. listwise deletion), this reduces the total number of cases, often quite considerably (as it only takes one missing value to exclude an entire case/person), whilst in the pairwise exclusion it means that exact comparison between sets of results may be impossible, given different sample sizes on each set of variables.

An alternative to the deletion method of excluding missing cases is the *imputation method* (e.g. single and multiple imputation). Imputation is a general term given to the many methods of trying to calculate what the missing values might be so that they can be included in the analysis, i.e. substituting missing values with plausible, calculated values (e.g. Rubin, 1987). This is beyond the scope of the present book, but it rests on techniques for making educated guesses in calculating probabilities, though there is no guarantee that this is 100 per cent accurate.

In single imputation, substitution (e.g. of means) takes place, for example entering the sample mean for the missing value. However, this reduces the chance of having true variability in the distributions and may compromise some statistical calculations such as correlational analysis. Multiple imputation methods rely on modelling the data, with maximum likelihood estimation and regression models used to predict and estimate missing values. Imputation can be used if the missing data are systematically missing, but they risk underestimating standard errors and are reliant on the robustness of the model being used. SPSS can calculate and enter missing values.

For more on missing values, we refer readers to Enders (2010), Carpenter and Kenward (2013), Little and Rubin (2014) and Raghunathan (2015).

40.2 Frequencies, percentages and crosstabulations

Frequencies and percentages

In descriptive statistics much is made of visual techniques of data presentation. Hence frequencies, percentages and forms of graphical presentation are often used. Many graphical forms of data presentation are available in software packages, including:

- frequency and percentage tables;
- bar charts (for nominal, ordinal and discrete data);
- histograms (for continuous – interval and ratio – data);
- line graphs;
- pie charts;
- high and low charts;
- scatterplots;
- stem and leaf displays;
- boxplots (box and whisker plots).

With most of these forms of data display there are various permutations of the ways in which data are displayed within the type of chart or graph chosen. Whilst graphs and charts may look appealing and have the benefit of visual immediacy, often they tell the reader no more than could be seen in a simple table of figures, and figures take up less space in a report and often carry more information. Pie charts, bar charts and histograms are particularly prone to this problem, and the data could be placed more succinctly into tables. Clearly the issue of fitness for audience is important here: some readers may find charts more accessible and comprehensible than tables of figures. Other charts and graphs can add greater value than tables, for example, line graphs, boxplots and scatterplots with regression lines, and these are helpful. Here are some guides on usage:

- bar charts are useful for presenting categorical and discrete data, highest and lowest (see Figure 40.1);
- avoid using a third dimension (e.g. depth) in a bar chart or histogram when it is unnecessary; a third dimension must provide additional information;
- histograms are useful for presenting continuous data;
- line graphs (single lines: one variable; or many lines: many variables) are useful for showing trends, particularly in continuous data, for one or more variables over time;
- pie charts and bar charts are useful for showing proportions;
- inter-dependence and relatedness can be shown through crosstabulations (discussed below);

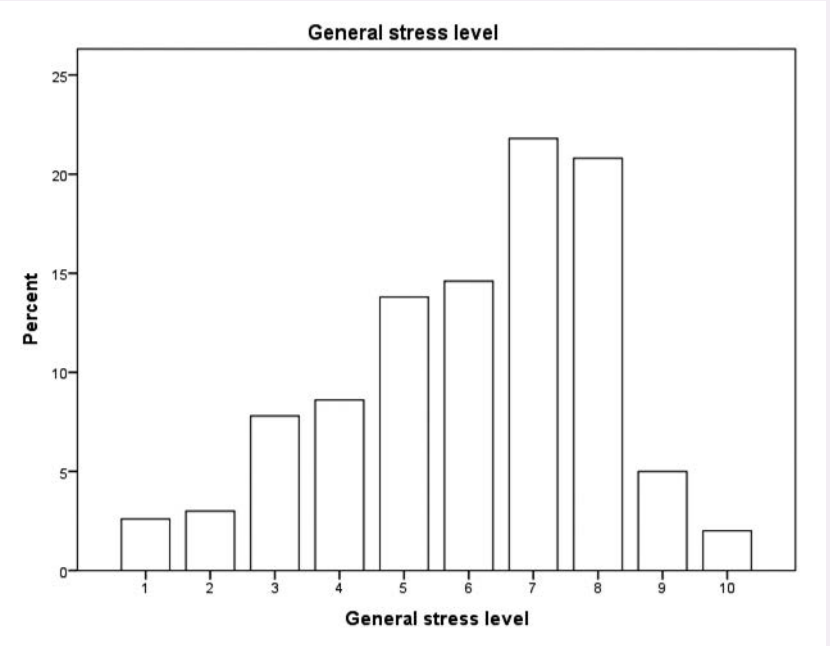


FIGURE 40.1 Bar chart of distribution of discrete stress levels amongst teachers (SPSS output)

- boxplots are useful for showing the distribution of values for several variables in a single chart, together with their range and medians (see Figure 40.2);
- stacked bar charts are useful for showing the frequencies or percentages of different groups within a specific variable for two or more variables in the same chart;
- scatterplots are useful for showing the relationship between two variables (see Figure 40.3).

With regard to bar charts, Figure 40.1 presents an example of teachers' reported stress levels.

In compiling Figure 40, 1,500 teachers reported their stress levels (from 1 to 10), and the bar chart sets out the percentages of respondents in each category. There are several points to observe here:

- the data are not normally distributed: they are negatively skewed (a long tail down to the left) and more respondents voted in the categories 7–10 (some 50 per cent) than in the categories 1–4 (22 per cent), i.e. more teachers were at the higher than the lower levels of stress'
- over 40 per cent of respondents voted in the categories 7 and 8'
- most of the results clustered around categories 5–8 (71 per cent).

The bar chart gives a powerful visual message but it does not present exact percentages. Compare this to Table 40.1 which lacks the visual impact but actually provides more detail. Researchers will need to decide which is fitter for purpose and audience.

With regard to boxplots, Figure 40.2 is rich in detail. It reports the mathematics test scores (marks out of 10)

TABLE 40.1 FREQUENCIES AND PERCENTAGES OF GENERAL STRESS LEVEL OF TEACHERS

Stress level	*Frequency*	*Percent*
1	13	2.6
2	15	3.0
3	39	7.8
4	43	8.6
5	69	13.8
6	73	14.6
7	109	21.8
8	104	20.8
9	25	5.0
10	10	2.0
Total	500	100.0

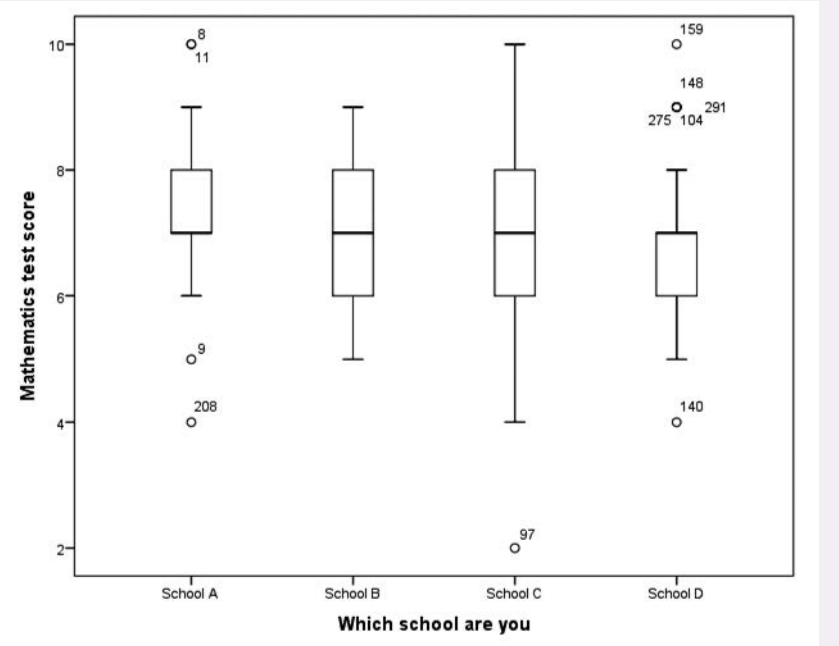


FIGURE 40.2 Boxplot of mathematics test scores in four schools (SPSS output)

for four schools, with 500 students in total, 125 from each school. There are several points to note from Figure 40.2:

- Each rectangular box contains the middle 50 per cent of cases for each school (the second and third quartile: a quartile is 25 per cent). We can see that School A and School D have small boxes, i.e. the middle 50% of cases (students) are in a narrow range of marks, whilst School B and School C have larger boxes, indicating that their middle 50 per cent of cases (students) are in a wider range of marks.
- the lines above and below each box (the 'whiskers' in this 'box and whisker' chart) indicate the range of the highest and lowest scores excluding outliers (some versions of the 'whiskers' have different definitions). Here we can see that Schools A, B and D have short whiskers (i.e. a narrow range), whilst School C has long whiskers, indicating a wider range of marks.
- the thick horizontal line in each box is the median value (the score of the middle case/person). As can be seen, in this instance the median value is the same for each school (the mark of 7 out of 10), but the distributions around that median are different. For example, in School A the median is at the bottom of the box, whereas in School D it is at the top, i.e. in School A the box (middle 50 per cent) contains marks of 7 and 8, whereas for School D the box (the middle 50 per cent) contains marks of 6 and 7, i.e. they are generally lower than the marks from School A.
- The small circles with numbers are outliers, i.e. those cases (students) whose scores are more than 1.5 times outside the interquartile range (discussed later in this chapter). The numbers next to these small circles are the case numbers in the SPSS file, so that the researcher can easily trace the exact outlier case (person) in the data set (each case has a number). School A has four outliers; School B has none; School C has one; and School D has six. Researchers will need to decide whether to remove or retain outliers. Outliers are true scores, but they may skew the mean and standard deviation (discussed below).

With regard to the scatterplot, Figure 40.3 shows the scores of students on a final university examination in relation to the number of hours per week that student spent on private study. Figure 40.3 contains several pieces of information:

- The small circles indicate the cases (the students), and the line from the lower left to the upper right is

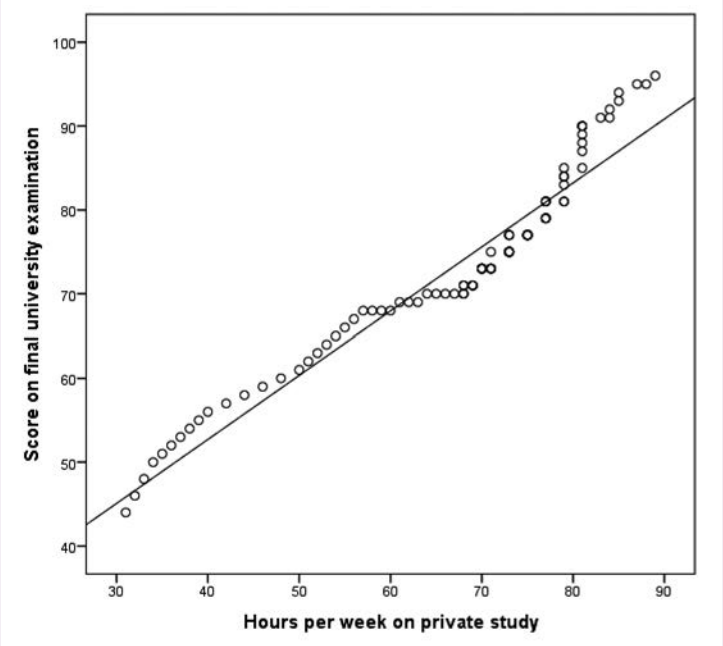


FIGURE 40.3 Scatter-plot with line of best fit (SPSS output)

the line of best fit of the scatter of the data. Many cases are close to the line of best fit, and some are a little further away from it; that is usual. The line of best fit is acceptable here, given that the cases are close to the line; in cases where the data are not close to the line of best fit, it may not be so useful. The line of best fit is a straight line, i.e. clearly there is a linear relationship between the two variables here.

- The scales of each axis do not start at zero, i.e. there were no students who studied for fewer than 30 hours each week, and there were no students who scored lower than 43 marks out of 100 (SPSS automatically calculates and presents the scale of marks to be used for each axis, and this can be edited at will).
- The more hours per week the students spent on private study, the higher was their university examination mark.
- The researcher can use this chart to predict scores, for example, a student who studies for fifty hours a week is likely to score 61 per cent in the examination, and a student who studies for seventy hours a week is likely to score 76 per cent on the examination.

At a simple level one can present data in terms of a table of frequencies and percentages (a piece of datum about a course evaluation), as shown in Table 40.2. From this table we can tell that:

a 191 people completed the item;
b most respondents thought that the course was 'a little' too hard (with a clear modal score of 98, i.e. 51.3 per cent); the modal score is that category or score which is given by the highest number of respondents;
c the results were skewed, with only 10.5 per cent being in the categories 'quite a lot' and 'a very great deal';

TABLE 40.2 FREQUENCIES AND PERCENTAGES FOR A COURSE EVALUATION (SPSS OUTPUT)

The course was too hard

		Frequency	Percentage
Valid	not at all	24	12.6
	very little	49	25.7
	a little	98	51.3
	quite a lot	16	8.4
	a very great deal	4	2.1
	Total	191	100.0

d more people thought that the course was 'not at all too hard' than thought that the course was 'quite a lot' or 'a very great deal' too hard;
e overall the course appears to have been slightly too difficult but not much more.

Crosstabulations

Let us imagine that we wished to explore further this piece of datum from Table 40.2. We may wish to discover, for example, the voting on this item by males and females. This can be presented in a simple crosstabulation, following the convention of placing the nominal data (male and female) in rows and the ordinal data (the five-point scale) in the columns (or independent variables as row data and dependent variables as column data). A crosstabulation is simply a presentational device, whereby one variable is presented in relation to another, with the relevant data inserted into each cell (automatically generated by software packages, such as SPSS) (Table 40.3).

Table 40.3 shows that, of the total sample, nearly three times more females (38.2 per cent) than males (13.1 per cent) thought that the course was 'a little' too hard, between two-thirds and three-quarters more females (19.9 per cent) than males (5.8 per cent) thought that the course was a 'very little' too hard, and around three times more males (1.6 per cent) than females (0.5 per cent) thought that the course was 'a very great deal' too hard. However, one also has to observe that the size of the two sub-samples was uneven. Around three-quarters of the sample was female (73.8%) and around one-quarter (26.2 per cent) was male.

There are two ways to overcome the problem of uneven sub-sample sizes. One is to adjust the sample, in this case by multiplying up the sub-sample of males by an exact figure in order to make the two sub-samples the same size (141/50=2.82). Another way is to examine the data by each row rather than by the overall totals, i.e. to examine the proportion of males voting such-and-such, and, separately, the proportion of females voting for the same categories of the variable, thus, as shown in Table 40.4.

If you think that these two calculations and re-calculations are complicated or difficult (overall-percentaged totals and row-percentaged totals), then be reassured: many software packages, for example, SPSS (the example used here) do this at one keystroke.

In Table 40.4 one can observe that:

- there was consistency in the voting by males and females in terms of the categories 'a little' and 'quite a lot';
- more males (6 per cent) than females (0.7 per cent) thought that the course was 'a very great deal' too hard;
- a slightly higher percentage of females (91.1 per cent: {12.1%+27%+52%}) than males (86 per cent: {14%+22%+50%}) indicated, overall, that the course was not too hard;
- the overall pattern of voting by males and females was similar, i.e. for both males and females the strong to weak categories in terms of voting percentages were identical.

We suggest that Table 40.4 is more helpful than Table 40.3, as, by including the row percentages, it renders fairer the comparison between the two groups: males and females. Further, we suggest that it is usually preferable to give *both* the actual frequencies and percentages, but to make the comparisons by percentages. We say this, because it is important for the reader to know the actual numbers used. For example, in Table 40.3, if we were simply to give the percentage of males voting that the

TABLE 40.3 CROSSTABULATION BY TOTALS (SPSS OUTPUT)

sex * The course was too hard: crosstabulation

		the course was too hard					
		not at all	very little	a little	quite a lot	a very great deal	Total
male	Count	7	11	25	4	3	50
	% of Total	3.7%	5.8%	13.1%	2.1%	1.6%	26.2%
female	Count	17	38	73	12	1	141
	% of Total	8.9%	19.9%	38.2%	6.3%	.5%	73.8%
Total	Count	24	49	98	16	4	191
	% of Total	12.6%	25.7%	51.3%	8.4%	2.1%	100.0%

TABLE 40.4 CROSSTABULATION BY ROW TOTALS (SPSS OUTPUT)

Sex * The course was too hard: crosstabulation

		the course was too hard					
		not at all	very little	a little	quite a lot	a very great deal	Total
male	Count	7	11	25	4	3	50
	% within sex	14.0%	22.0%	50%	8.0%	6.0%	100%
female	Count	17	38	73	12	1	141
	% within sex	12.1%	27.0%	52%	8.5%	.7%	100%
Total	Count	24	49	98	16	4	191
	% within sex	12.6%	25.7%	51%	8.4%	2.1%	100%

course was a 'very great deal' too hard (1.6 per cent), as course planners we might worry about this. However, when we realize that 1.6 per cent is actually only three out of 141 people then we might be less worried. Had the 1.6 per cent represented, say, fifty people of a sample, then this would have given us cause for concern. Percentages on their own can mask the real numbers, and the reader needs to know the real numbers.

Researchers can comment on particular cells of a crosstabulated matrix in order to draw attention to certain factors (e.g. the very high 52 per cent in comparison to its neighbour 8.5 per cent in the voting of females in Table 40.4). It is also useful, on occasions, to combine data from more than one cell, as we have done in the example above. For example, if we combine the data from the males in the categories 'quite a lot' and 'a very great deal' (8%+6%=14%) we can observe that, not only is this equal to the category 'not at all', but it also contains fewer cases than any of the other single categories for the males, i.e. the combined category shows that the voting for the problem of the course being too difficult is still very slight.

Combining categories can be useful in showing the general trends or tendencies in the data. For example, in Tables 40.2–40.4, combining 'not at all', 'very little' and 'a little', all of these measures indicate that it is only a very small problem of the course being too hard, i.e. generally speaking the course was not too hard.

Combining categories can also be useful in rating scales of agreement to disagreement. For example, consider the following results in relation to for example, a survey of 200 people on a particular item (Table 40.5).

There are several ways of interpreting Table 40.5, for example: (a) more people 'strongly agreed' (20 per cent) than 'strongly disagreed' (15 per cent); (b) the modal score was for the central neutral category (a central tendency) of 'neither agree nor disagree'. However one can go further. If one wishes to ascertain an overall indication of disagreement and agreement then adding together the two disagreement categories yields 35 per cent (15%+20%) and adding together the two agreement categories yields 30 per cent (10%+20%), i.e. there was more disagreement than agreement, despite the fact that more respondents 'strongly agreed' than 'strongly disagreed', i.e. the *strength* of agreement and disagreement has been lost. Adding together the two disagreement and agreement categories gives us a general rather than a detailed picture; this may be useful for our purposes. However, if we do this then we also have to draw attention to the fact that the total of the two disagreement categories (35 per cent) is the same as the total in the category 'neither agree nor disagree', in which case one could suggest that the modal category of 'neither agree nor disagree' has been superseded by bi-modality, with disagreement being one modal score and 'neither agree nor disagree' being the other.

Combining categories can be useful, though it is not without its problems; for example let us consider three tables (Tables 40.6–40.8). The first presents the overall results of an imaginary course evaluation, in which

TABLE 40.5 RATING SCALE OF AGREEMENT AND DISAGREEMENT

Strongly disagree	*Disagree*	*Neither agree nor disagree*	*Agree*	*Strongly agree*
30 15%	40 20%	70 35%	20 10%	40 20%

TABLE 40.6 SATISFACTION WITH A COURSE

	Satisfaction with course			
	Low (1–3)	*Medium (4–5)*	*High (6–7)*	*Total*
Male	60 (41.4%)	70 (48.3%)	15 (10.3%)	145 (100%)
Female	35 (43.7%)	15 (18.8%)	30 (37.5%)	80 (100%)
Total	95 (42.2%)	85 (37.8%)	45 (20%)	225 (100%)

TABLE 40.7 COMBINED CATEGORIES OF RATING SCALES

	Satisfaction with course		
	Low (1–5)	*High (6–7)*	*Total*
Male	130 (89.7%)	15 (10.3)	145 (100%)
Female	50 (62.5%)	30 (37.5%)	80 (100%)
Total	180 (76.1%)	45 (23.9%)	225 (100%)
Difference	+27.2%	–27.2%	

TABLE 40.8 REPRESENTING COMBINED CATEGORIES OF RATING SCALES

	Satisfaction with course		
	Low (1–3)	*High (4–7)*	*Total*
Male	60 (41.4%)	85 (58.6%)	145 (100%)
Female	35 (43.7%)	45 (56.3%)	80 (100%)
Total	95 (42.6%)	130 (57.4%)	225 (100%)
Difference	–2.1%	+1.9%	

three levels of satisfaction have been registered (low, medium, high) (Table 40.6).

Here one can observe that the modal category (the category with the most votes) is 'low' (95 votes: 42.2 per cent) and the lowest category is 'high' (45 votes: 20 per cent), i.e. overall the respondents are dissatisfied with the course. The females seem to be more satisfied with the course than the males, if the category 'high' is used as an indicator, and the males seem to be more moderately satisfied with the course than the females.

However, if one combines categories (low and medium) then a different story could be told (Table 40.7). By looking at the percentages in Table 40.7, it appears that the females are more satisfied with the course overall than males, and that the males are more dissatisfied with the course than females.

However, if one were to combine categories differently (medium and high) then a different story could be told (Table 40.8). By looking at the percentages in Table 40.8, it appears that there is not much difference between the males and the females, and that both males and females are highly satisfied with the course. At issue here are dangers in combining categories (collapsing tables), and we advocate great caution in doing this. Sometimes it can provide greater clarity, and sometimes it can distort the picture. In the example it is wiser to keep with the original table rather than collapsing it into fewer categories.

Crosstabulations for categorical data can be bivariate (two variables presented), for example Table 40.9, in which two forms of primary students (Primary 3 and Primary 4) are asked how interesting they find a course. The rows are the nominal, categorical variable and the columns are the values of the ordinal variable. This is a commonplace organization.

Let us give another example. Suppose that we wished to examine the views of parents from socially advantaged and disadvantaged backgrounds of primary school children on traditional school examinations (in favour/against), using simple dichotomous variables (two values only in each variable). Table 40.10 presents the results. It shows us clearly that parents from socially advantaged backgrounds are more in favour of formal, written public examinations than those from socially disadvantaged backgrounds.

Additionally, a trivariate crosstabulation can be constructed (and SPSS enables researchers to do this), with three variables included. In the example here, let us say that we are interested in their socio-economic status (socially advantaged/socially disadvantaged) and their philosophies of education (traditionalist/child-centred). Our results appear in Table 40.11.

The results here are almost the reverse of Table 40.10: now the socially advantaged are more likely than socially disadvantaged parents to favour forms of assessment other than formal, written public examinations, and socially disadvantaged parents are more likely than socially advantaged parents to favour formal, written public examinations. The educational philosophies of each of the two groups (socially advantaged and socially disadvantaged) have dramatically altered the scenario. A trivariate analysis can give greater subtlety to the data and their analysis. (Introducing a third variable can be used as a control variable, and we address this in the discussion of correlations.)

Box 40.1 provides the SPSS command sequence for crosstabulations.

TABLE 40.9 A BIVARIATE CROSSTABULATION (SPSS OUTPUT)

form * the contents are interesting Crosstabulation

			the contents are interesting				Total
			strongly agree	agree	no comment	disagree	
form	Primary 3	Count	12	7	1	2	22
		% within form	54.5%	31.8%	4.5%	9.1%	100.0%
	Primary 4	Count	35	18	11		64
		% within form	54.7%	28.1%	17.2%		100.0%
	Total	Count	47	25	12	2	86
		% within form	54.7%	29.1%	14.0%	2.3%	100.0%

TABLE 40.10 A BIVARIATE ANALYSIS OF PARENTS' VIEWS ON PUBLIC EXAMINATIONS

Formal, written public examinations	*Acceptability of formal, written public examinations*	
	Socially advantaged	*Socially disadvantaged*
In favour	70%	35%
Against	30%	65%
Total per cent	100%	100%

TABLE 40.11 A TRIVARIATE CROSSTABULATION

Formal, written public examinations	*Acceptability of formal, written public examinations*			
	Traditionalist		*Progressivist/child-centred*	
	Socially advantaged	*Socially disadvantaged*	*Socially advantaged*	*Socially disadvantaged*
In favour	65%	70%	35%	20%
Against	35%	30%	65%	80%
Total per cent	100%	100%	100%	100%

BOX 40.1 SPSS COMMAND SEQUENCE FOR CROSSTABULATIONS

The SPSS command sequence for Crosstabulations is: 'Analyze' → 'Descriptive Statistics' → 'Crosstabs' → Enter the row variable in the 'Rows' box and the column variable in the 'Columns' box → Click the 'Cells' box, which opens a new window → In the 'Percentages' area, check the 'Total' box' → Click 'Continue' → Click 'OK'.

40.3 Measures of central tendency and dispersal

Central tendency: means, modes and medians

The central tendency of a set of scores investigates how they cluster round the middle of the set of scores, or where the majority of scores are located. For categorical data the measure of central tendency is the mode: that score which is given by the most people, that score which has the highest frequency (there can be more than one mode: if there are two clear modal scores then this is termed 'bi-modal'; if there are three then this is termed 'tri-modal'). For continuous data (e.g. ratio data), in addition to the mode, the researcher can calculate the mean (the average score) and the median (the middle score, e.g. of the middle person): half of the scores fall above it and half below it (the median is also sometimes used for ordinal data). If there is an even number of observations then the median is the average of the two middle scores. We cannot calculate the median score for nominal data, as the data have to be rank-ordered from the lowest to the highest in terms of the quantity of the variable under discussion. Measures of central tendency are used with univariate data, and indicate the typical score (the mode), the middle score (the median) and the average score (the mean).

As a general rule, the mean is a useful statistic if the data are not skewed (i.e. if they are not bunched at one end of a curve of distribution or do not conform to the normal curve of distribution) or if there are no outliers that may be exerting a disproportionate effect (including a high standard deviation, see below). One has to recall that the mean, as a statistical calculation only, can sometimes yield some strange results, for example fractions of a person!

The median is useful for ordinal data, but, to be meaningful, there have to be many scores rather than just a few. The median overcomes the problem of outliers, and hence is useful for skewed results or those with a wide dispersal (i.e. high standard deviation). The modal score is useful for all scales of data, particularly nominal and ordinal data, i.e. discrete and categorical data, rather than continuous data, and it is unaffected by outliers, though it is not strong if there are many values and many scores which occur with similar frequency.

The standard deviation

Are scores widely dispersed around the mean, do they cluster close to the mean, or are they at some distance from the mean? The measures used to determine this are measures of dispersal. If we have interval and ratio data then, in addition to the modal score and crosstabulations, we can calculate the mean (the average) and the standard deviation. The standard deviation is the average distance that each score is from the mean, i.e. the average difference between each score and the mean, and how much the scores, as a group, deviate from the mean. It is a standardized measure of dispersal. For small samples (fewer than thirty scores), or for samples rather than populations, it is calculated as:

$$SD = \sqrt{\frac{\sum d^2}{N-1}}$$

where

d^2 = the deviation of the score from the mean (average), squared

$\sum$ = the sum of

N = the number of cases

For populations rather than samples, it is calculated as:

$$SD = \sqrt{\frac{\sum d^2}{N}}$$

A low standard deviation indicates that the scores cluster together, whilst a high standard deviation indicates that the scores are widely dispersed. This is calculated automatically by software packages such as SPSS, at the click of a single button.

Let us imagine that we have the test scores for 1,000 students, on a test that was marked out of 10

TABLE 40.12 DISTRIBUTION OF TEST SCORES (SPSS OUTPUT)

Test scores

		Frequency	Valid Percent
Valid	2	1	.1
	3	223	22.3
	4	276	27.6
	5	32	3.2
	6	69	6.9
	7	149	14.9
	8	185	18.5
	9	39	3.9
	10	26	2.6
	Total	1000	100.0

(Table 40.12). Here we can calculate that the average score was 5.48. We can also calculate the standard deviation. In the example here the standard deviation in the example of scores was 2.134. What do these tell us? Firstly, it suggests that the average mark was not very high (5.48). Secondly, it tells us that there was quite a variation in the scores. Thirdly, the scores were unevenly spread, indeed there was a large cluster of scores around the categories of 3 and 4, and another large cluster of scores around the categories 7 and 8. This is where a line graph could be useful in representing the scores, as it shows two peaks clearly, as shown in Figure 40.4.

It is important to report the standard deviation. For example, let us consider the following. Look at these three sets of numbers:

(1)	1	2	3	4	20	mean = 6
(2)	1	2	6	10	11	mean = 6
(3)	5	6	6	6	7	mean = 6

If we were to plot the points in (1), (2) and (3) onto three separate graphs we would see very different results (Figures 40.5–40.7). Figure 40.5 shows the mean being heavily affected by the single score of 20 (an 'outlier': an extreme score a long way from the others); in fact all the other four scores are some distance below the mean. The score of 20 is exerting a disproportionate effect on the data and on the mean, raising it. Some statistical packages (e.g. SPSS) can exclude outliers. If the data are widely spread then it may be more suitable not to use the median score; SPSS performs this automatically at the click of a key.

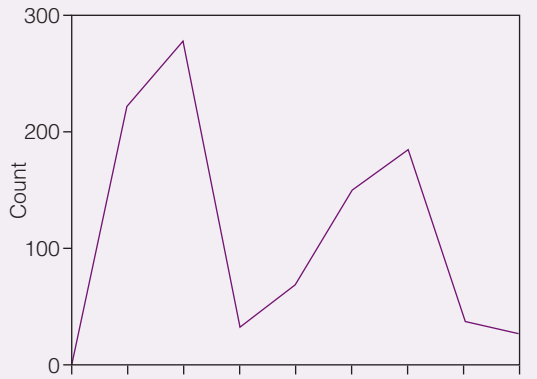


FIGURE 40.4 A line graph of test scores

Figure 40.6 shows one score actually on the mean but the remainder some distance away from it. The scores are widely dispersed and the shape of the graph is flat (a platykurtic distribution).

Figure 40.7 shows the scores clustering very tightly around the mean, with a very peaked shape to the graph (a leptokurtic distribution).

The point is this: it is not enough simply to calculate and report the mean; for a fuller picture of the data we need to look at the dispersal of scores. For this we can use the standard deviation, though the standard deviation is susceptible to the disproportionate effects of outliers. Some scores will be widely dispersed (the first graph), others will be evenly dispersed (the second graph) and others will be bunched together (the third graph). A high standard deviation will indicate a wide dispersal of scores, a low standard deviation will indicate clustering or bunching together of scores.

The range

A second way of measuring dispersal is to calculate the range, which is the difference between the lowest (minimum) score and the highest (maximum) score in a set of scores. This incorporates extreme scores, and is susceptible to the distorting effect of outliers: a wide range may be found if there are outliers, and if these outliers are removed then the range may be much reduced. Further, the range tells the researcher nothing about the distributions of scores within the range.

The interquartile range

Another measure of dispersal is the interquartile range. If we arrange a set of scores in order, from the lowest to the highest, then we can divide that set of scores into four equal parts: the lowest quarter (quartile) that contains the lowest quarter of all the scores, the lower-middle quartile, the upper-middle quartile, and the highest quarter (quartile) that contains the highest quarter of the scores. The interquartile range is the difference between the first quartile and the third quartile, or more precisely the difference between the 25th and the 75th percentile, i.e. the middle 50 per cent of scores (the second and third quartiles). This, thereby, ignores extreme scores and, unlike the simple range, does not change significantly if the researcher adds some scores that are some distance away from the average. For example, let us imagine that we have a set of test scores thus, ordered into quartiles:

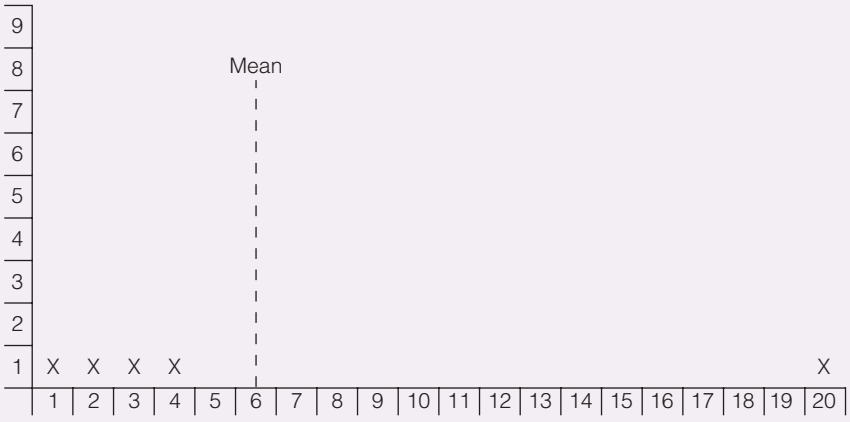


FIGURE 40.5 Distribution around a mean with an outlier

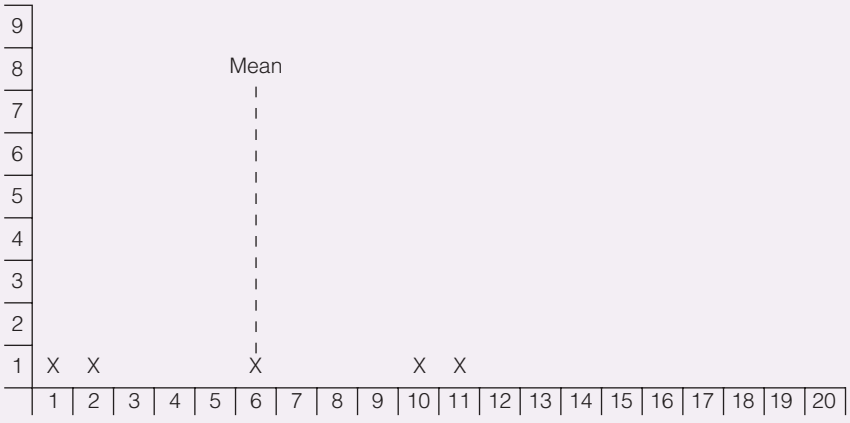


FIGURE 40.6 A platykurtic distribution of scores

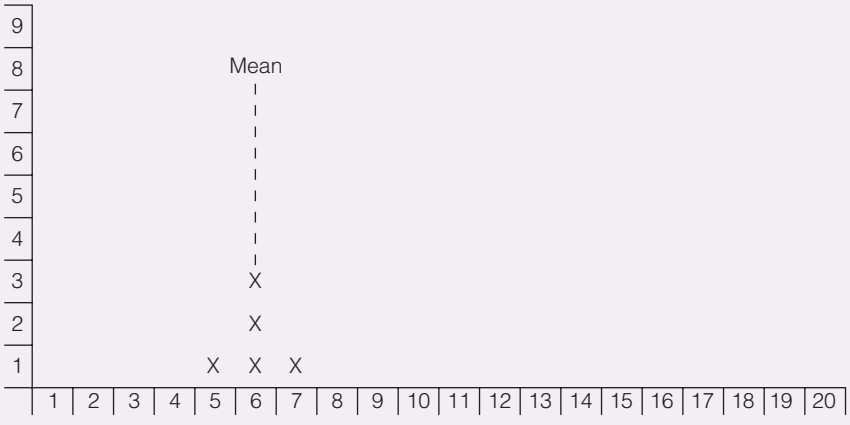


FIGURE 40.7 A leptokurtic distribution of scores

FIRST QUARTILE	*SECOND QUARTILE*	*THIRD QUARTILE*	*FOURTH QUARTILE*
40	50	**65**	83
41	55	70	86
43	58	75	90
47	63	77	93

The interquartile range is 47–65 (emboldened figures), which is 18. There are other ways of calculating the interquartile range (e.g. the difference between the medians of the first and third quartiles, the difference between the median of the lower half of the data and the median of the upper half of the data), and the reader may wish to explore these.

Though there are several ways of calculating dispersal, by far the most common is the standard deviation.

The SPSS command sequence for frequencies, standard deviations, Standard Error, skewness and kurtosis, range, means, modes and median is provided in Box 40.2.

40.4 Taking stock

What we do with simple frequencies and descriptive data depends on the scales of data that we have (nominal, ordinal, interval and ratio) (NB interval and ratio data are combined as 'scale' data in SPSS). For all four scales we can calculate frequencies and percentages, and present these in a variety of forms. We can also calculate the mode and present crosstabulations, both bivariate and trivariate crosstabulations. We can combine categories and collapse tables into smaller tables, providing that the sensitivity of the original data has not been unfairly lost. We can calculate the median score, which is particularly useful if the data are spread widely (with high standard deviations) or if there are outliers. For interval and ratio data we can also calculate the mean and the standard deviation; the mean yields an average and the standard deviation indicates the range of dispersal of scores around that average, i.e. to see whether the data are widely dispersed (e.g. a platykurtic distribution), or close together with a distinct peak (a leptokurtic distribution). We can use other measures of dispersal such as the range and the interquartile range. In examining frequencies and percentages, researchers can investigate whether the data are skewed, i.e. over-represented at one end of a scale and under-represented at the other end. A positive skew has a long tail at the positive end and the majority of the data at the negative end, and a negative skew has a long tail at the negative end and the majority of the data at the positive end.

40.5 Correlations and measures of association

Much educational research is concerned with establishing relationships between variables. We may wish to know, for example, how achievement is related to social class background; whether an association exists between the number of years spent in full-time education and subsequent income; whether there is a link between personality and achievement. What, for example, is the relationship, if any, between membership of a public library and social class status? Is there a relationship between social class background and placement in different strata of the secondary school curriculum, or between gender and success/failure in 'first-time' driving test results?

There are several simple measures of association readily available to researchers to help them test these sorts of relationships. We have selected the most widely used ones and set them out in Table 40.13. Of these, the two most commonly used correlations are the Spearman rank order correlation for ordinal data and the Pearson product moment correlation for interval and ratio data, and we advise readers to use these as the main kinds of correlation statistics. At this point it is pertinent to say a few words about some of the terms used in Table 40.13 to describe the nature of variables. Cohen and Holliday (1982, 1996) provide worked examples of the appropriate use and limitations of the

BOX 40.2 SPSS COMMAND SEQUENCE FOR DESCRIPTIVE STATISTICS

The SPSS command sequence for frequencies, standard deviations, Standard Error, skewness and kurtosis, range, means, modes and median is: 'Analyze' → 'Descriptive Statistics' → 'Frequencies' → Send over to the 'Variables' box the variables in which you are interested → Click the 'Statistics' box which will open a new window → Check the measures of central tendency which you wish to calculate (mean, mode, median). Check the measures of dispersion which you wish to calculate (standard deviation, variance, range, maximum, minimum, Standard Error of the mean). Check the measures of distribution which you wish to calculate (skewness, kurtosis) → Click 'Continue' → Click the 'Charts' box → Check the kind of chart which you wish to have (histogram, pie, bar) and the chart values (frequencies, percentages) → Click 'Continue' → Click 'OK'.

TABLE 40.13 COMMON MEASURES OF RELATIONSHIP

Measure	*Nature of variables*	*Comment*
Spearman's rho	Two ordinal variables	Relationship linear
Pearson product moment, *r*	Two continuous variables; interval or ratio scale	Relationship linear
Rank order or Kendall's tau	Two continuous variables; ordinal scale	
Correlation ratio, η (eta)	One variable continuous, other either continuous or discrete	Relationship non-linear
Intraclass	One variable continuous; other discrete; interval or ratio scale	Purpose: to determine within-group similarity
Biserial, r_{bis} Point biserial, $r_{pt\ bis}$	One variable continuous; other (a) continuous but dichotomized. r_{bis} or (b) true dichotomy, $r_{pt\ bis}$	Index of item discrimination (used in item analysis)
Phi coefficient, φ	Two true dichotomies; nominal or ordinal series	
Partial correlation $r_{12.3}$	Three or more continuous variables	Purpose: to determine relationship between two variables, with effect of third held constant
Multiple correlation $r_{1.234}$	Three or more continuous variables	Purpose: to predict one variable from a linear weighted combination of two or more independent variables
Kendall's coefficient of concordance (W)	Three or more continuous variables; ordinal series	Purpose: to determine the degree of (say, inter-rater) agreement

Source: Mouly (1978)

correlational techniques outlined in Table 40.13, together with other measures of association such as Kruskal's *gamma*, Somer's *d* and Guttman's *lambda*.

Look at the words used at the top of the table to explain the nature of variables in connection with the measure called the Pearson product moment, *r*. The variables are 'continuous' and at the 'interval' or the 'ratio' scale of measurement. A continuous variable is one that, theoretically at least, can take any value between two points on a scale. Weight, for example, is a continuous variable; so too is time, so also is height. Weight, time and height can take on any number of possible values between nought and infinity, the feasibility of measuring them across such a range being limited only by the variability of suitable measuring instruments.

Turning again to Table 40.13, we read in connection with the third measure shown there (rank order or Kendall's *tau*) that the two continuous variables are at the ordinal scale of measurement.

The variables involved in connection with the *phi* coefficient measure of association in Table 40.13 are described as 'true dichotomies' and at the nominal scale of measurement. Truly dichotomous variables (such as sex or driving test result) can take only two values (male or female; pass or fail).

To conclude our explanation of terminology, readers should note the use of the term 'discrete variable' in the description of the fourth correlation ratio (*eta*) in Table 40.13. We said earlier that a continuous variable can take on any value between two points on a scale. A discrete variable, however, can only take on numerals or values that are specific points on a scale. The number of players in a football team is a discrete variable. It is usually eleven; it could be fewer than eleven, but it could never be seven-and-a-quarter!

Box 40.3 provides the SPSS command sequence for correlations.

BOX 40.3 SPSS COMMAND SEQUENCE FOR CORRELATIONS

In SPSS the command sequence for correlations is: 'Analyze' → 'Correlate' → 'Bivariate' → Send to the box marked 'Variables' the variables which you wish to correlate. Check the box with the correlation statistic which you wish to calculate (Pearson, Spearman). Check the radio button for the test of significance (one-tailed, two-tailed) → Click 'OK'.

The percentage difference

The percentage difference is a simple asymmetric measure of association. An *asymmetric* measure is a measure of *one-way association*, that is to say, it estimates the extent to which one phenomenon implies the other but not *vice versa*. Gender, as we shall see shortly, may imply driving test success or failure. The association could never be the other way round. Measures which are concerned with the extent to which two phenomena imply each other are referred to as *symmetric* measures. Table 40.14 reports the percentage of public library members by their social class origin. What can we discover from the data set out in Table 40.14? By comparing percentages in different columns of the same row, we can see that 49 per cent more middle class persons are members of public libraries than working-class persons. By comparing percentages in different rows of the same columns we can see that 72 per cent more middle class persons are members rather than non-members. The data suggest an association between the social class status of individuals and their membership of public libraries.

A second way of making use of the data in Table 40.14 involves the computing of a *percentage ratio* (%R). Look, for example, at the data in the second row of Table 40.14. By dividing 63 by 14 (%R=4.5) we can say that four-and-a-half times as many working-class persons are not members of public libraries as are middle-class persons.

The *percentage difference* ranges from 0 per cent when there is complete independence between two phenomena to 100 per cent when there is complete association in the direction being examined. It is straightforward to calculate and simple to understand. Notice, however, that the percentage difference as we have defined it can only be employed when there are only two categories in the variable along which we percentage and only two categories in the variable in which we compare. In SPSS, using the 'Crosstabs' command can yield percentages, and we indicate this in the website manual that accompanies this volume.

TABLE 40.14 PERCENTAGE OF PUBLIC LIBRARY MEMBERS BY THEIR SOCIAL CLASS ORIGIN

Public library membership	*Social class status*	
	Middle class	*Working class*
Member	86	37
Non-member	14	63
Total	100	100

In connection with this issue, on the accompanying website we discuss the *phi* coefficient, the correlation coefficient tetrachoric r (r_t), the contingency coefficient C and combining independent significance tests of partial relations.

Explaining correlations

In our discussion of the principal correlational techniques shown in Table 40.13, three are of special interest to us and these form the basis of much of the rest of the chapter. They are the Pearson product moment correlation coefficient, multiple correlation and partial correlation.

Correlational techniques are generally intended to answer three questions about two variables or two sets of data. First, 'Is there a relationship between the two variables (or sets of data)?' If the answer to this question is 'yes', then two other questions follow: 'What is the direction of the relationship?' and 'What is the magnitude of the association?'

Relationship in this context refers to any tendency for the two variables (or sets of data) to vary consistently. Pearson's product moment coefficient of correlation, one of the best-known measures of association, is a statistical value ranging from –1.0 to +1.0 and expresses this relationship in quantitative form. The coefficient is represented by the symbol *r*.

Where the two variables (or sets of data) fluctuate in the same direction, i.e. as one increases so does the other, or as one decreases so does the other, a positive relationship is said to exist. Correlations reflecting this pattern are prefaced with a plus sign to indicate the positive nature of the relationship. Thus +1.0 indicates perfect positive correlation between two factors, as with the radius and diameter of a circle, and +0.80 a high positive correlation, as between academic achievement and intelligence, for example. Where the sign has been omitted, a plus sign is assumed.

A negative correlation or relationship, on the other hand, is found when an increase in one variable is accompanied by a decrease in the other variable. Negative correlations are prefaced with a minus sign. Thus –1.0 would represent perfect negative correlation, as between the number of errors children make on a spelling test and their score on the test, and –0.30 a low negative correlation, as, say, between absenteeism and intelligence. There is no other meaning to the signs used; they indicate nothing more than which pattern holds for any two variables (or sets of data).

Researchers are interested in the magnitude of an obtained correlation as well as in its direction.

Correlational procedures have been developed so that no relationship whatever between two variables is represented by zero (or 0.00), as between body weight and intelligence, possibly. This means that a person's performance on one variable is totally unrelated to her performance on a second variable. If she is high on one, for example, she is just as likely to be high or low on the other. Perfect correlations of +1.00 or –1.00 are rarely found and, as we shall see, most coefficients of correlation in social research are around +0.50 or less. The correlation coefficient may be seen, then, as an indication of the predictability of one variable given the other: it is an indication of covariation. The relationship between two variables can be examined visually by plotting the paired measurements on a graph, with each pair of observations being represented by a point. The resulting arrangement of points is a 'scatterplot' and enables us to assess graphically the degree of relationship between the characteristics being measured (see Figure 40.3). Figure 40.8 gives some examples of scatterplots in educational research.

Whilst correlations are widely used in research, and they are straightforward to calculate and to interpret, the researcher must be aware of four caveats in undertaking correlational analysis:

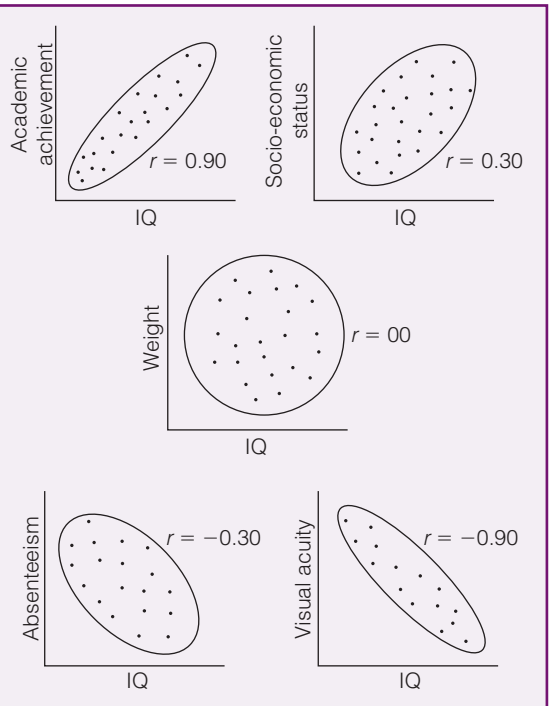


FIGURE 40.8 Correlation scatterplots
Source: Tuckman (1972)

i Do not assume that correlations imply causal relationships (i.e. simply because having large hands appears to correlate with having large feet does not imply that having large hands causes one to have large feet).
ii Be alert to a Type I error: not supporting the null hypothesis when it is in fact true (a false positive).
iii Be alert to a Type II error: supporting the null hypothesis when it is in fact not true (a false negative).
iv Given the problems of statistical significance set out in Chapter 39, it must be accompanied by an indication of effect size (the coefficient of correlation).

In SPSS a typical print-out of a correlation coefficient is given in Table 40.15. In this fictitious example, using 1,000 cases, there are four points to note:

i The cells of data to the right of the cells containing the figure 1 are the same as the cells to the left of the cells containing the figure 1, i.e. there is a mirror image and researchers can decide whether to look at only the variables to the right of the cell with the figure 1 (the perfect correlation, since it is one variable being correlated with itself), or to look at the cells to the left of the figure 1.
ii In each cell where one variable is correlated with a different variable there are three figures: the top figure gives the correlation coefficient, the middle figure gives the significance level and the lowest figure gives the sample size.
iii SPSS marks with an asterisk those correlations which are statistically significant.
iv All the correlations are positive, since there are no negative coefficients given.

These tables give us the magnitude of the correlation (the coefficient), the direction of the correlation (positive and negative) and the significance level. The correlation coefficient is the effect size. The significance level is calculated automatically by SPSS, based on the coefficient and the sample size: the greater the sample size, the lower the coefficient of correlation has to be in order to be statistically significant, and, by contrast, the smaller the sample size, the greater the coefficient of correlation has to be in order to be statistically significant (see also Chapter 39).

In reporting correlations one has to report the statistic used, the coefficient, the direction of the correlation (positive or negative) and the significance level (if considered appropriate). For example, one could write:

TABLE 40.15 A PEARSON PRODUCT MOMENT CORRELATION (SPSS OUTPUT)

		The attention given to teaching and learning at the school	How well students apply themselves to learning	Discussion and review by educators of the quality of teaching, learning and classroom practice
The attention given to teaching and learning at the school	Pearson Correlation	1	.060	.066*
	Sig. (2-tailed)	.	.058	.036
	N	1000	1000	1000
How well students apply themselves to learning	Pearson Correlation	.060	1	.585**
	Sig. (2-tailed)	.058	.	.000
	N	1000	1000	1000
Discussion and review by educators of the quality of teaching, learning and classroom practice	Pearson Correlation	.066*	.585**	1
	Sig. (2-tailed)	.036	.000	.
	N	1000	1000	1000

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

> Using the Pearson product moment correlation, a large effect size (correlation coefficient $r=0.87$) and a statistically significant correlation ($\rho=0.035$) were found between students' attendance at school and their examination performance. Those students who attended school the most tended to have the best examination performance, and those who attended the least tended to have the lowest examination performance.

Alternatively, there may be occasions when it is important to report when a correlation has *not* been found, for example:

> There was a very small effect size (correlation coefficient $r=0.16$) and no statistically significant correlation found ($\rho=0.43$) between the amount of time spent on homework and examination performance.

In both of these examples of reporting, exact significance levels have been given, assuming that SPSS has calculated these. An alternative way of reporting the significance levels (if appropriate) are: $\rho < 0.05$; $\rho < 0.01$; $\rho < 0.001$; $\rho=0.05$; $\rho=0.01$, $\rho=0.001$. In the case of statistical significance not having been found one could report this as $\rho > 0.05$ or $\rho=$N.S. (not significant).

Curvilinearity

The correlations discussed so far have assumed linearity, that is, the more we have of one property, the more (or less) we have of another property, in a direct positive or negative relationship. A straight line can be drawn through the points on the scatterplots (a line of best fit). However, linearity cannot always be assumed. Consider the case, for example, of stress: a little stress might enhance performance positively ('setting the adrenalin running'), whereas too much stress might lead to a downturn in performance. Where stress enhances performance there is a positive correlation, but when stress debilitates performance there is a negative correlation. The result is not a straight line of correlation (indicating linearity) but a curved line (indicating curvilinearity). This can be shown graphically (Figure 40.9). It is assumed here, for the purposes of the example, that muscular strength can be measured on a single scale. It is clear from the graph

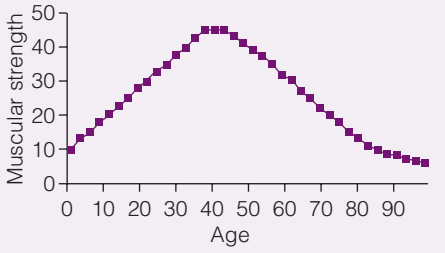


FIGURE 40.9 A line diagram to indicate curvilinearity

that muscular strength increases from birth until fifty years, and thereafter it declines as muscles degenerate. There is a positive correlation between age and muscular strength on the left-hand side of the graph and a negative correlation on the right-hand side of the graph, i.e. a curvilinear correlation can be observed.

Hopkins *et al*. (1996, p. 92) provide another example of curvilinearity: room temperature and comfort. Raising the temperature a little can make for greater comfort – a positive correlation – whilst raising it too greatly can make for discomfort: a negative correlation. Many correlational statistics assume linearity (e.g. the Pearson product moment correlation). However, rather than using correlational statistics arbitrarily or blindly, the researcher will need to consider whether, in fact, linearity is a reasonable assumption to make, or whether a curvilinear relationship is more appropriate (in which case more sophisticated statistics will be needed, e.g. η (*eta*)) (Glass and Hopkins, 1996, section 8.7; Cohen and Holliday, 1996, p. 84; Fowler *et al*., 2000) or mathematical procedures will need to be applied to transform non-linear relations into linear relations. Examples of curvilinear relationships might include:

- pressure from the principal and teacher performance;
- pressure from the teacher and student achievement;
- degree of challenge and student achievement;
- assertiveness and success;
- age and muscular strength;
- age and physical control;
- age and concentration;
- age and sociability;
- age and cognitive abilities.

Hopkins *et al*. (1996, p. 92) suggest that poorly constructed tests can give the appearance of curvilinearity if the test is too easy (a 'ceiling effect' where most students score highly) or if it is too difficult, but that this curvilinearity is, in fact, spurious, as the test does not demonstrate sufficient item difficulty or discriminability (see Chapter 27).

In planning correlational research, then, attention will need to be given to whether linearity or curvilinearity is to be assumed.

Coefficients of correlation

The coefficient of correlation tells us about the relations between two variables. Other measures exist, however, which allow us to specify relationships when more than two variables are involved. These are known as measures of 'multiple correlation' and 'partial correlation'.

Multiple correlation measures indicate the degree of association between three or more variables simultaneously. We may want to know, for example, the degree of association between delinquency, social class background and leisure facilities. Or we may be interested in finding out the relationship between academic achievement, intelligence and neuroticism. Multiple correlation, or 'regression' as it is sometimes called, indicates the degree of association between *n* variables. It is related not only to the correlations of the independent variable with the dependent variables, but also to the intercorrelations between the dependent variables.

Partial correlation aims at establishing the degree of association between two variables after the influence of a third has been controlled or partialled out. Guilford and Fruchter (1973) define a partial correlation between two variables as one which nullifies the effects of a third variable (or a number of variables) on the variables being correlated.

Consider, for example, the relationship between (a) success in basketball and (b) previous experience in the game. Suppose, also, that the presence of a third factor (c) – the height of the players – was known to have an important influence on the other two factors (a) and (b). The use of partial correlation techniques would enable a measure of the two primary variables (a) and (b) to be achieved, freed from the influence of the secondary variable (c).

Correlational analysis is simple and involves collecting two or more scores on the same group of subjects and computing correlation coefficients. Many useful studies have been based on this simple design. Those involving more complex relationships, however, utilize multiple and partial correlations in order to provide a clearer picture of the relationships being investigated.

One final point: it is important to stress again that correlations refer to measures of association and do not necessarily indicate causal relationships between variables. Correlation does not imply cause.

Interpreting the correlation coefficient

Once a correlation coefficient has been calculated, there remains the problem of interpreting it. A question often asked here is how large should the coefficient be for it to be meaningful. The question may be approached in three ways: by examining the strength of the relationship; by examining the square of the correlation coefficient; and, if the researcher adheres to the value of significance testing (see Chapter 39), by examining the statistical significance of the relationship (though significance testing does not indicate the magnitude of the correlation, only the likelihood of a chance relationship).

Inspection of the numerical value of a correlation coefficient yields clear indication of the strength of the relationship between the variables in question. Low or near zero values indicate weak relationships, while those nearer to +1 or –1 suggest stronger relationships. Imagine, for instance, that a measure of a teacher's success in the classroom after five years in the profession is correlated with her final school experience grade as a student and that it was found that $r=+0.19$. Suppose now that her score on classroom success is correlated with a measure of need for professional achievement and that this yielded a correlation of 0.65. It could be concluded that there is a stronger relationship between success and professional achievement scores than between success and final student grade.

Where a correlation coefficient has been derived from a sample and one wishes to use it as a basis for inference about the parent population, the statistical significance of the obtained correlation can be considered. Statistical significance, when applied to a correlation coefficient, indicates whether or not the correlation is different from zero at a given level of confidence. As we have seen earlier, a statistically significant correlation is indicative of an actual relationship rather than one due entirely to chance, even though its assumption of the null hypothesis is questionable (Chapter 39). The level of statistical significance of a correlation is determined to a great extent by the number of cases upon which the correlation is based: the greater the number of cases, the smaller the correlation need be to be statistically significant at a given level of confidence.

The second approach to interpreting a coefficient is provided by examining the square of the coefficient of correlation, r^2. This shows the proportion of variance in one variable that can be attributed to its linear relationship with the second variable. In other words, it indicates the amount the two variables have in common. If, for example, two variables A and B have a correlation of 0.50, then $(0.50)^2$ or 0.25 of the variation shown by the B scores can be attributed to the tendency of B to vary linearly with A. Figure 40.10 shows graphically the common variance between reading grade and arithmetic grade having a correlation of 0.65.

Third, many exploratory relationship studies are interpreted with reference to their statistical significance, whereas prediction studies depend for their efficacy on the strength of the correlation coefficients (the effect size). Here correlation coefficients need to be considerably higher than those found in exploratory relationship studies and for this reason rarely invoke the concept of statistical significance.

There are three cautions to be borne in mind when one is interpreting a correlation coefficient. First, a coefficient is a simple number and must not be interpreted as a percentage. A correlation of 0.50, for instance, does not mean a 50 per cent relationship between the variables. Further, a correlation of 0.50 does not indicate twice as much relationship as that shown by a correlation of 0.25. A correlation of 0.50 actually indicates more than twice the relationship shown by a correlation of 0.25. In fact, as coefficients approach +1 or –1, a difference in the absolute values of the coefficients becomes more important than the same numerical difference between lower correlations would be.

Second, a correlation does not necessarily imply a cause-and-effect relationship between two factors, as previously indicated. It should not therefore be interpreted as meaning that one factor is causing the scores on the other to be as they are. There are invariably other factors influencing both variables under consideration. Suspected cause-and-effect relationships would have to be confirmed by other kinds of study.

Third, a correlation coefficient is not to be interpreted in any absolute sense. A correlational value for a given sample of a population may not necessarily be the same as that found in another sample from the same

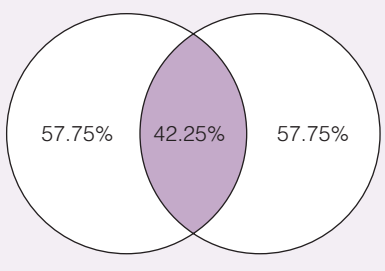


FIGURE 40.10 Visualization of correlation of 0.65 between reading grade and arithmetic grade

Source: Fox (1969)

population. Many factors influence the value of a given correlation coefficient and if researchers wish to extrapolate to the populations from which they drew their samples they will then have to consider testing the significance of the correlation or the sampling strategy used.

We now offer some general guidelines for interpreting correlation coefficients. They are based on Borg's (1963) analysis and assume that the correlations relate to 100 or more subjects.

Correlations ranging from 0.20 to 0.35

Correlations within this range show only very slight relationship between variables although they may be statistically significant. A correlation of 0.20 shows that only 4 per cent ($\{0.20 \times 0.20\} \times 100$) of the variance is common to the two measures. Whereas correlations at this level may have limited meaning in exploratory relationship research, they are of no value in either individual or group prediction studies.

Correlations ranging from 0.35 to 0.65

Within this range, correlations are statistically significant beyond the 1 per cent level. When correlations are around 0.40, crude group prediction may be possible. As Borg notes, correlations within this range are useful, however, when combined with other correlations in a multiple regression equation. Combining several correlations in this range in some cases can yield individual predictions that are correct within an acceptable margin of error. Correlations at this level used singly are of little use for individual prediction because they yield only a few more correct predictions than could be accomplished by guessing or by using some chance selection procedure.

Correlations ranging from 0.65 to 0.85

Correlations within this range make possible group predictions that are accurate enough for most purposes. Nearer the top of the range, group predictions can be made very accurately, usually predicting the proportion of successful candidates in selection problems within a very small margin of error. Near the top of this correlation range, individual predictions can be made that are considerably more accurate than would occur if no such selection procedures were used.

Correlations over 0.85

Correlations as high as this indicate a close relationship between the two variables correlated. A correlation of 0.85 indicates that the measure used for prediction has about 72 per cent variance in common with the performance being predicted. Prediction studies in education very rarely yield correlations this high. When correlations at this level are obtained, however, they are very useful for either individual or group prediction.

40.6 Partial correlations

Many researchers wish to control for the effects of other variables. As we discussed in Chapter 6, controlling for the effects of variables means holding them constant whilst manipulating other variables. Let us imagine that we examine the scores of 500 students on a mathematics test to see if there is any relationship between their test scores (marks out of 100) and how easy they find mathematics (scored out of 100). We conduct a Pearson correlation and find a large positive correlation (correlation coefficient of 0.738, indicated in Table 40.16), which is statistically significant $\rho = 0.000$.

However, we want to know if this positive correlation holds true when other variables are controlled, in this case the variable 'how interested are you in mathematics?' (Table 40.17), so we conduct a partial correlation. Partial correlations enable the researcher to control for a third variable, i.e. to see the correlation between two variables of interest once the effects of a third variable have been removed, hence rendering more accurate the relationship between the two variables of interest. Indeed SPSS will enable more than one control variable to be inserted. Partialling can rule out special or specific relationships that do not hold true when variables have been controlled. Using SPSS, partial correlations can be calculated straightforwardly at the touch of a key.

In our example we have 'mathematics test score', 'how easy do you find mathematics?', and, as the control variable, 'how interested are you in mathematics?'. This time we see that, from the SPSS output in Table 40.17, when we control for the third variable ('how interested are you in mathematics?'), the correlation coefficient between 'mathematics test score' and 'how easy do you find mathematics?' drops massively from 0.738 to 0.071, an extremely low correlation or no real correlation at all, which, this time, is not statistically significant ($\rho = 0.112$). In other words, when we control for the third variable, the degree of association between the two initial variables reduces; the third variable exerts a considerable effect on the strength of the relationship between the initial two variables.

Imagine, this time, that instead of controlling for 'how interested are you in mathematics?' we control for a different third variable: 'how much do you like mathematics?' (Table 40.18). This time the correlation coefficient between the original two variables 'mathematics test score' and 'how easy do you find

TABLE 40.16 CORRELATION BETWEEN SCORE ON MATHEMATICS TEST AND HOW EASY THE STUDENTS FIND MATHEMATICS (SPSS OUTPUT)

Correlations

		Mathematics test score	How easy do you find mathematics?
Mathematics test score	Pearson Correlation Sig. (2-tailed) N	1 500	0.738** 0.000 500
How easy do you find mathematics?	Pearson Correlation Sig. (2-tailed) N	0.738** 500	1 500

Note
** Correlation is significant at the 0.01 level (2-tailed).

TABLE 40.17 CORRELATION BETWEEN SCORE ON MATHEMATICS TEST AND HOW EASY THE STUDENTS FIND MATHEMATICS, CONTROLLING FOR STUDENTS' INTEREST IN MATHEMATICS (SPSS OUTPUT)

Correlations

Control variables			Mathematics test score	How easy do you find mathematics?
How interested are you in mathematics?	Mathmatics test score	Correlation Significance (2-tailed) df	1.000 0	0.071 0.112 497
	How easy do you find mathematics?	Correlation Significance (2-tailed) df	0.071 0.112 497	1.000 0

TABLE 40.18 CORRELATION BETWEEN SCORE ON MATHEMATICS TEST AND HOW EASY THE STUDENTS FIND MATHEMATICS, CONTROLLING FOR STUDENTS' LIKING OF MATHEMATICS (SPSS OUTPUT)

Correlations

Control variables			Mathematics test score	How easy do you find mathematics?
How much do you like mathematics?	Mathmatics test score	Correlation Significance (2-tailed) df	1.000 0	0.711 0.000 497
	How easy do you find mathematics?	Correlation Significance (2-tailed) df	0.711 0.000 497	1.000 0

mathematics?' has changed only a very small amount, from 0.738 to 0.711, and is still statistically significant ($\rho = 0.000$). Here the third variable has made almost no difference to the strength of the correlation between the two original variables, i.e. controlling for 'how much do you like mathematics?' has had very little effect on the strength of the relationship between the two original variables.

Partial correlation, then, enables relationships to be calculated after controlling for one or more variables.

In SPSS the command sequence for partial correlations is provided in Box 40.4.

40.7 Reliability

Correlation can be used in reliability testing. We need to know how reliable the items in our instrument are for data collection. Reliability in quantitative analysis takes two main forms, both of which are measures of internal consistency: the split-half technique and the alpha coefficient. Both calculate a coefficient of reliability that can lie between 0 and 1. The formula for calculating the Spearman-Brown split-half reliability (discussed in Chapter 14) is:

$$r = \frac{2r}{1+r}$$

where r = the actual correlation between the halves of the instrument (this requires the instrument to be able to be divided into two matched halves in terms of content and difficulty). So, for example, if the correlation coefficient between the two halves is 0.85 then the formula would be worked out thus:

$$r = \frac{2(0.85)}{1+0.85} = \frac{1.70}{1.85} = 0.919$$

Here the split-half reliability coefficient is 0.919, which is very high. SPSS automatically calculates split-half reliability at the click of a button.

An alternative calculation of reliability as internal consistency can be found in Cronbach's alpha, frequently referred to simply as the alpha coefficient of reliability. The Cronbach alpha provides a coefficient of inter-item correlations by calculating the average of all possible split-half reliability coefficients. It is a measure of the internal consistency among the *items* (not the people/cases) and is used for multi-item scales. SPSS calculates Cronbach's alpha at the click of a key. The formula for alpha is:

$$alpha = \frac{nr_{ii}}{1+(n-1)r_{ii}}$$

where n = the number of items in the test or survey (e.g. questionnaire) and r_{ii} = the average of all the inter-item correlations. Let us imagine that the number of items in the survey is 10, and that the average correlation is 0.738. The alpha correlation can be calculated thus:

$$alpha = \frac{nr_{ii}}{1+(n-1)r_{ii}} = \frac{10(0.738)}{1+(10-1)0.739} = \frac{7.38}{7.64} = 0.97$$

This yields an alpha coefficient of 0.97, which is very high. For the split-half coefficient and the alpha coefficient the following guidelines can be used:

>0.90 very highly reliable
0.80–0.90 highly reliable
0.70–0.79 reliable
0.60–0.69 marginally/minimally reliable
<0.60 unacceptably low reliability

Bryman and Cramer (1990, p. 71) suggest that the reliability level is acceptable at 0.8, though others suggest that it is acceptable if it is 0.67 or above.

If the researcher is using SPSS then there is a function which enables items to be discovered that might be exerting a negative influence on the Cronbach alpha. Table 40.19 provides an example of this, which indicates in the final column what the Cronbach alpha would be if any of the items were to be removed as being unreliable.

In the example, the overall alpha is given as 0.642, but it can be seen that if the item 'How much do you

BOX 40.4 SPSS COMMAND SEQUENCE FOR PARTIAL CORRELATIONS

In SPSS the command sequence for partial correlations is: 'Analyze' → 'Correlate' → 'Partial' → Send to the box marked 'Variables' the variables which you wish to correlate. Send to the box marked 'Controlling for' the control variable(s) that you wish to include. Check the radio button for the test of significance (one-tailed, two-tailed) → Click 'OK'.

feel that you treat colleagues as impersonal objects?' were removed, then the overall reliability would rise to 0.658. The researcher, then, may wish to remove that item; this is particularly true where the researcher is conducting a pilot to see which items are reliable and which are not.

In SPSS the command sequence for the Cronbach alpha is provided in Box 40.5.

TABLE 40.19 IDENTIFYING UNRELIABLE ITEMS IN CRONBACH'S ALPHA (SPSS OUTPUT)

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
How hard do you feel you are working in your job?	17.71	51.379	.272	.642
How much do you feel exhausted by the end of the workday?	17.99	43.902	.472	.553
How much do you feel that you cannot cope with your job any longer?	21.02	38.863	.567	.495
How much do you feel that you treat colleagues as impersonal objects?	23.52	50.959	.242	.658
How much do you feel that working with colleagues all day is really a strain for you?	22.08	43.084	.436	.569

Overall alpha: .642

BOX 40.5 SPSS COMMAND SEQUENCE FOR RELIABILITY CALCULATION

In SPSS the command sequence for the Cronbach alpha is: 'Analyze' → 'Scale' → 'Reliability Analysis' → Send to the box marked 'Items' the variables which you wish to include → Click the box marked 'Statistics'. This opens a new window. In the area marked 'Descriptives for', check the box marked 'Scale if item deleted' → Click 'Continue' → Click 'OK'.

Companion Website

The companion website to the book provides additional material, data files and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Inferential statistics

CHAPTER 41

Difference tests

The previous chapter introduced descriptive statistics. This chapter moves to inferential statistics, those statistics that enable researchers to make inferences about the wider population. Here we introduce difference tests and how they can be conducted. The chapter proceeds thus:

- measures of difference between groups
- the t-test (a test of difference for parametric data)
- Analysis of Variance (a test of difference for parametric data)
- the chi-square test for non-parametric data (a test of difference and a test of goodness of fit)
- degrees of freedom (a statistic that is used in calculating statistical significance in considering difference tests)
- the Mann-Whitney and Wilcoxon tests (tests of difference for non-parametric data)
- the Kruskal-Wallis and Friedman tests (tests of difference for non-parametric data)

These statistics constitute powerful tools in the arsenal of statistics for analysing numerical data. We give several worked examples for clarification, and take the novice reader by the hand through these.

41.1 Measures of difference between groups

Researchers will sometimes be interested to investigate whether there are differences between two or more groups of sub-samples, answering questions such as: 'Is there a statistically significant difference between the amount of homework done by boys and girls, and if so, how much?'; 'Is there a statistically significant difference between test scores from four similarly mixed-ability classes studying the same syllabus, and if so, how much?'; 'Does school A differ statistically significantly from school B in the stress level of its sixth form students, and if so, by how much?'. Such questions require measures of difference. This section introduces measures of difference and how to calculate difference. The process can commence with the null hypothesis, stating that 'there is no statistically significant difference between the two groups', or 'there is no statistically significant difference between the four groups', and, if this is not supported, then the alternative hypothesis is supported, namely, there is a statistically significant difference between the two (or more) groups'. Then it can proceed to tests of magnitude of difference (effect size). We discuss difference tests for parametric and non-parametric data. We discussed statistical significance and its limitations in Chapter 39, and we suggested very strongly that it should be accompanied by measures of effect size. Statistical significance suggests whether a result occurs by chance (though, as Chapter 39 suggests, this is, in fact, open to question), whilst effect size calculates how much of a difference there is.

Before going very far one has to ascertain:

- the kind of data with which one is working (parametric or non-parametric), as this affects the choice of statistic used;
- the number of groups being compared (e.g. two or more groups), to discover whether there is a difference between them;
- whether the groups are related or independent. Independent groups are entirely unrelated to each other, for example, males and females completing an examination; related groups might be the same group voting on two or more variables or the same group voting at two different points in time (e.g. a pre-test and a post-test).

Statistics are usually divided into those which work with parametric or non-parametric data, those which measure differences between two groups and those which measure differences between more than two groups, and whether the groups are related or independent. Decisions on these matters affect the choice of statistics used. Our discussion proceeds thus: *first* we look at a simple difference test for two groups using parametric data, which is the t-test (Section 41.2). *Second* we look at differences between three or more groups using parametric data: Analysis of Variance (ANOVA)

with *post hoc* tests (the Tukey test and the Games-Howell test) (Section 41.3). *Third* we look at a test of difference for categorical data (the chi-square test) (Section 41.4). *Fourth*, we introduce the 'degrees of freedom' (Section 41.5). *Fifth* we look at differences between two groups using non-parametric data (the Mann-Whitney and Wilcoxon tests) (Section 41.6); *sixth* we look at differences between three or more groups using non-parametric data (the Kruskal-Wallis and the Friedman tests) (Section 41.7). As in previous examples, we use SPSS to illustrate our points.

41.2 The t-test

The t-test is used to discover whether there are statistically significant differences between the means of two groups or for the same group under two conditions, drawn from random samples with a normal distribution and using parametric data in the dependent variable. It is used to compare the means of two groups randomly assigned, for example on a pre-test and a post-test in an experiment, or for the same group under two conditions. The t-test operates under certain assumptions, which we list here as 'safety checks', i.e. to see if it is safe to proceed with the use of the test.

Safety checks for using the t-test

The t-test requires several 'safety checks':

- parametric continuous data with the dependent variable at interval or ratio level;
- random sampling;
- normal distribution of the data (though large samples often overcome this);
- equality of variance (similarity/equality of variance in each group: 'homogeneity of variance'), though the Levene test can overcome problems here, and SPSS calculates this automatically.

If these safety requirements are not met then the researcher should use a non-parametric difference test (e.g. Mann-Whitney U test; the Wilcoxon test), even if the data are interval or ratio.

Conducting the-test

The t-test has two variants: the t-test for independent samples and the t-test for related (or 'paired') samples. The former assumes that the two groups are unrelated to each other; the latter assumes that it is the *same*, single group either voting on two variables or voting at two different points in time about the same variable or under two conditions. We will address the t-test for independent samples first. The t-test assumes that one variable is categorical (e.g. males and females) and one is a continuous variable (e.g. marks on a test). The formula calculates a statistic based on:

$$t = \frac{\text{Sample one mean} - \text{sample two mean}}{\text{Standard error of the difference in means}}$$

Let us imagine that we wish to discover whether, concerning how well learners are cared for, guided and supported, there is a statistically significant difference between (a) the leader/senior management team (SMT) of a group of randomly chosen schools and (b) the teachers. The data are ratio; the participants have awarded a mark out of ten for their response; and the higher the mark, the greater the care, guidance and support offered to the students. The t-test for two independent samples presents us with two tables in SPSS. First it provides the average (mean) of the voting for each group: 8.37 for the leaders/senior managers and 8.07 for the teachers, i.e. there is a difference of means between the two groups. Is this difference statistically significant, i.e. by chance or otherwise? Is the null hypothesis ('there is no statistically significant difference between the leaders/SMT and the teachers') supported or not supported? We commence with the null hypothesis ('there is no statistically significant difference between the two means') and then we set the level of significance (α) to use for supporting or not supporting the null hypotheses; for example we could say 'Let $\alpha = 0.05$'. Then the data are computed as in Table 41.1.

In running the t-test, SPSS gives us back what, at first glance, seems to be a morass of information in Table 41.1. Some of this is superfluous for our purposes here. We will concern ourselves with the most important pieces of data for introductory purposes here: the Levene test and the significance level for a two-tailed test (Sig. (2-tailed)) (Table 41.2).

The Levene test is a guide as to which row of the two to use ('equal variances assumed' and 'Equal variances not assumed'). Look at the column 'Sig.' in the Levene test (0.004). If the probability value is statistically significant (as in this case: 0.004) then variances are *unequal* and the researcher needs to use the second row of data ('Equal variances not assumed'); if the probability value is not significant ($\rho > 0.05$) then equal variances *are* assumed and she/he uses the first row of data ('Equal variances assumed'). Once she/he has decided which row to use then the Levene test has served its purpose and the researcher can move on. For our commentary here, the purpose of the Levene test is only there to determine which row to look at of the two presented.

Having discovered which row to follow, in our example it is the second row, we go along to the

TABLE 41.1 MEANS AND STANDARD DEVIATIONS FOR A t-TEST (SPSS OUTPUT)

Group Statistics

	who are you	N	Mean	Std. Deviation	Std. Error Mean
How well learners are cared for, guided and supported	leader/member of the SMT	347	8.37	2.085	.112
	teachers	653	8.07	2.462	.096

TABLE 41.2 THE LEVENE TEST FOR EQUALITY OF VARIANCES IN A t-TEST (SPSS OUTPUT)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
How well learners are cared for, guided and supported	Equal variances assumed	8.344	.004	1.92	998	.055	.30	.155	-.006	.603
	Equal variances not assumed			2.02	811.922	.044	.30	.148	.009	.589

column 'Sig. (2-tailed)'. This tells us that there is a statistically significant difference between the two groups – leaders/SMT and the teachers – because the significance level is 0.044 (i.e. $\rho < 0.05$). Hence we can say that the null hypothesis is not supported, that there is a statistically significant difference between the means of the two groups ($\rho = 0.044$), and that the mean of the leaders/SMT is statistically significantly higher (8.37) than the mean of the teachers (8.07), i.e. the leaders/SMT of the schools think more highly than the teachers in the schools that the learners are well cared for, guided and supported, and that this difference is not by chance.

Look at Table 41.2 again, and at the column 'Sig. (2-tailed)'. Had equal variances been assumed (i.e. if the Levene test had indicated that we should remain on the top row of data rather than the second row of data) then we would *not* have found a statistically significant difference between the two means ($\rho = 0.055$, i.e. $\rho > 0.05$). Hence it is sometimes important to know whether equal variances are to be assumed or not to be assumed.

In the example here we find that there *is* a statistically significant difference between the means of the two groups, i.e. the leaders/SMT do not share the same perception as the teachers that the learners are well cared for, guided and supported. Typically the leaders/SMT are more generous than the teachers, and this difference is not by chance. This is of research interest, for example, to discover the reasons for, and impact of, the differences of perception. It could be, for example, that the leaders/SMT have a much rosier picture of the situation than the teachers, and that the teachers – the ones who have to work with the students on a close daily basis – are more in touch with the students and know that there are problems, a matter to which the senior managers may be turning a blind eye.

In reporting the t-test here the following form of words can be used:

> The mean score of the leaders/SMT on the variable 'How well learners are cared for, guided and supported' ($M = 8.37$, $SD = 2.02$) is statistically significantly higher ($t = 1.92$, $df = 811.922$, two-tailed $\rho = 0.044$) than those of teachers on the same variable ($M = 8.07$, $SD = 2.462$).

In this example, significance testing suggests to the researcher whether or not the difference found is by chance alone; that's all. This may be of limited use to

the researcher who wants to know *how much* difference there is, and here we refer the reader to the tests of effect size addressed in Chapter 39, for example, Cohen's *d*. Effect size can be calculated straightforwardly with online calculators, and we indicate these in Chapter 39, alongside how to perform hand calculations of effect size.

Let us take a second example. Here the leaders/SMT and teachers are voting on 'the attention given to teaching and learning in the school', again awarding a mark out of ten, i.e. ratio data. The mean for the leaders/SMT is 5.53 and for the teachers it is 5.46. Are these means statistically significantly different (Tables 41.3 and 41.4) or is there a difference that is not due to chance alone?

If we examine the Levene test (Sig.) in Table 41.4 we find that equal variances *are* assumed ($\rho = 0.728$), i.e. we remain on the top row of the data output. Running along to the column headed 'Sig. (2-tailed)' we find that $\rho = 0.610$, i.e. there is no statistically significant difference between the means of the two groups, therefore the null hypothesis (there is no statistically significant difference between the means of the two groups) is supported. In other words, the difference between the means is by chance alone. This should not dismay the researcher; finding or *not* finding a statistically significant difference is of equal value in research: a win–win situation. Here, for example, one can say that there is a shared perception between the leaders/managers and the teachers on the attention given to teaching and learning in the school, even though the attention given is poor (means of 5.53 and 5.46 respectively). The fact that there is a shared perception – that both parties see the same problem in the same way – offers a positive prospect for development and a shared vision, i.e. even though the picture is poor, nevertheless it is perhaps more positive than if there were very widely different perceptions.

In reporting the t-test here the following form of words can be used:

> The mean score for the leaders/SMT on the variable 'the attention given to teaching and learning at the school' ($M = 5.53$, $SD = 2.114$) did not differ statistically significantly ($t = 0.510$, $df = 998$, two-tailed $\rho = 0.610$) from that of the teachers ($M = 5.46$, $SD = 2.145$).

In this example, too, significance testing suggests to the researcher whether or not the difference found is by

TABLE 41.3 A t-TEST FOR LEADERS AND TEACHERS (SPSS OUTPUT)

Group Statistics

	who are you	N	Mean	Std. Deviation	Std. Error Mean
The attention given to teaching and learning at the school	leader/member of the SMT	347	5.53	2.114	.113
	teachers	653	5.46	2.145	.084

TABLE 41.4 THE LEVENE TEST FOR EQUALITY OF VARIANCES BETWEEN LEADERS AND TEACHERS (SPSS OUTPUT)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
The attention given to teaching and learning at the school	Equal variances assumed	.121	.728	.510	998	.610	.07	.142	-.206	.351
	Equal variances not assumed			.513	714.630	.608	.07	.141	-.205	.350

chance alone; that's all. This may be of limited use to the researcher who wants to know *how much* difference there is, and, as before, we refer the reader to the tests of effect size and how to calculate it, set out in Chapter 39.

The t-test for independent examples is a very widely used statistic, and we support its correct use very strongly.

The t-test can also be used for a paired (related) sample, i.e. where the same group votes on two variables (e.g. liking for mathematics and music), or the same *group* is measured on two occasions (e.g. the pre-test and the post-test) or under two conditions (e.g. morning and evening), or the same *variable* is measured at two points in time (pre-test and post-test). Here two variables are paired, with marks awarded by the same group (Table 41.5).

One can look to see if the mean of the 1,000 respondents who voted on 'the attention given to teaching and learning in the school' (mean = 5.48) is statistically significantly different from the mean of the same group voting on the variable 'the quality of the lesson preparation' (mean = 7.17) (Table 41.6).

In Table 41.6 we can move directly to the final column ('Sig. (2-tailed)') where we find that $\rho = 0.000$, i.e. $\rho < 0.001$, telling us that the null hypothesis is not supported, and that there is a statistically significant difference between the two means (i.e. the result is not by chance alone), even though it is the same group that is awarding the marks.

The issue of testing the difference between two proportions is set out on the accompanying website.

Box 41.1 provides the SPSS command sequence for the t-test for independent samples.

Box 41.2 provides the SPSS command sequence for the t-test for related (paired) samples.

To complement statistical significance in difference measurement, calculating effect size can be conducted using 'partial eta squared' and Cohen's *d*. (In SPSS the command sequence is: 'Analyze' → 'General Linear Model' → 'Univariate' → 'Estimates of effect size'.) Eta squared is also used, and is the proportion of the

TABLE 41.5 MEANS AND STANDARD DEVIATIONS IN A PAIRED SAMPLES t-TEST (SPSS OUTPUT)

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	The attention given to teaching and learning at the school	5.48	1000	2.134	.067
	The quality of the lesson preparation	7.17	1000	1.226	.039

TABLE 41.6 THE PAIRED SAMPLES t-TEST (SPSS OUTPUT)

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	The attention given to teaching and learning at the school - The quality of the lesson preparation	-1.69	2.430	.077	-1.84	-1.54	-21.936	999	.000

BOX 41.1 SPSS COMMAND SEQUENCE FOR INDEPENDENT SAMPLES T-TEST

To run the t-test for independent samples in SPSS the command sequence is: 'Analyze' → 'Compare means' → 'Independent samples T Test' → Send the dependent variable to box 'Test variable' and the independent variable to the 'Grouping variable' box → Click 'Define groups' (which is activated when the 'Grouping variable' box contains the independent variable) and then type the number that you assigned to each of the two groups in the SPSS file (e.g. males '1' and females '2') → Click 'Continue' (which returns you to the original screen) → Click 'OK'.

BOX 41.2 SPSS COMMAND SEQUENCE FOR T-TEST FOR RELATED (PAIRED) SAMPLES

To run the t-test for related (paired) samples in SPSS it is important, first, for the researcher to define the single group to be observed under the two conditions. The single group might be, for example, only the males from a total sample of males and females. Here SPSS requires you to use the Select Cases function (the command sequence in SPSS is: 'Data' → 'Select Cases'). Then decide which radio button you wish to activate ('If the condition is satisfied'; 'random sample of cases'; 'Based on time or case range'; 'Use filter variable'); each of these open another box for further selection and instructions (cf. Pallant, 2016). Once you have selected the cases (NB if you do not use this function then the entire sample is used), the SPSS command sequence for the t-test is: 'Analyze' → 'Compare means' → 'Paired samples T Test' → Click the first variable in which you are interested and send it to the 'Variable 1' box, and then click the second variable in which you are interested and send it to the 'Variable '2 box → Click 'OK'.

total variance that can be attributed to a particular effect; partial eta squared is the proportion of the effect plus error variance that can be attributed to a particular effect. Partial eta squared and Cohen's *d* are the preferred measures here.

41.3 Analysis of Variance

The t-test is useful for examining differences between *two* groups of respondents, or the same group on either two variables or two occasions, using parametric data from a random sample and assuming that each datum value is independent of the others. However, in much educational research we may wish to investigate differences between *more than two* groups. For example, we may wish to look at the examination results of four regions or four kinds of schools. In this case the t-test will not suit our purposes, and we must turn to Analysis of Variance. Analysis of Variance (ANOVA) can be used with three or more groups and is premised on the same assumptions as t-tests, and the research should conduct these 'safety checks' to ensure that it is appropriate to use ANOVA.

Safety checks for using ANOVA

Analysis of Variance (ANOVA) requires:

- continuous parametric data;
- random sampling;
- normal distribution of the data (though large samples often overcome this);
- homogeneity (equality) of variances (though the Levene test can identify problems here, and SPSS can offer the Brown-Forsythe and Welch tests to overcome the problem here, discussed below).

There are several kinds of Analysis of Variance; here we introduce only the three most widely used versions: the one-way Analysis of Variance, the two-way Analysis of Variance and Multiple Analysis of Variance (MANOVA). Analysis of Variance, like the t-test, assumes that the independent variable(s) is/are categorical (e.g. teachers, students, parents, governors) and one is a continuous variable (e.g. marks on a test). It calculates the F ratio, given as:

$$\text{F ratio} = \frac{\text{Between-groups variance}}{\text{Within-groups variance}}$$

ANOVA calculates the means for all the groups and then it calculates the average of these means. For each group separately it calculates the total deviation of each individual's score from the mean of the group (within-groups variation). Finally it calculates the deviation of each group mean from the grand mean (between-groups variation).

One-way Analysis of Variance

Let us imagine that we have four types of school:

- rural primary
- rural secondary
- urban primary
- urban secondary.

Let us imagine further that all of the schools in these categories have taken the same standardized test of mathematics, and the results have been given as a percentage thus, as shown in Table 41.7.

Table 41.7 gives us the means, standard deviations, standard error, confidence intervals, and the minimum and maximum marks for each group. At this stage we are only interested in the means:

rural primary: mean = 59.85%
rural secondary: mean = 60.44%
urban primary: mean = 50.64%
urban secondary: mean = 51.70%

Are these means statistically significantly different, i.e. real differences or by chance alone? Analysis of Variance (ANOVA) calculates statistical significance. We commence with the null hypothesis ('there is no statistically significant difference between the four means') and then we set the level of significance (α) to use for supporting or not supporting the null hypothesis; for example, we could say 'Let $\alpha = 0.05$'. SPSS performs the calculation (Table 41.8).

Table 41.8 tells us that, for three degrees of freedom (df), the F-ratio is 8.976. The F-ratio is the *between*-group mean square (variance) divided by the *within*-group mean square (variance), i.e.:

$$F = \frac{\text{Between-group variance}}{\text{Within-group variance}} = \frac{3981.040}{443.514} = 8.976$$

Looking at the final column ('Sig.'), ANOVA tell us that there is a statistically significant difference between the means ($\rho = 0.000$). This does *not* mean that all the means are statistically significantly different from each other, but that some are. For example, it may be that the means

TABLE 41.7 DESCRIPTIVE STATISTICS FOR ANALYSIS OF VARIANCE (SPSS OUTPUT)

Descriptives

Standardised Mathematics scores (percentages)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Rural primary	134	59.85	21.061	1.819	56.25	63.45	30	100
Rural secondary	136	60.44	19.470	1.669	57.14	63.74	30	100
Urban primary	141	50.64	22.463	1.892	46.90	54.38	30	100
Urban secondary	194	51.70	21.077	1.513	48.72	54.69	30	100
Total	605	55.22	21.473	.873	53.51	56.94	30	100

TABLE 41.8 SPSS OUTPUT FOR ONE-WAY ANALYSIS OF VARIANCE (SPSS OUTPUT)

ANOVA

Standardised Mathematics scores (percentages)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	11943.119	3	3981.040	8.976	.000
Within Groups	266551.8	601	443.514		
Total	278494.9	604			

for the rural primary and rural secondary schools (59.85 per cent and 60.44 per cent respectively) are not statistically significantly different, and that the means for the urban primary schools and urban secondary schools (50.64 per cent and 51.70 per cent respectively) are not statistically significantly different. However, it could be that there is a statistically significant difference between the scores of the rural (primary and secondary) and the urban (primary and secondary) schools. How can we find out which groups are statistically significantly different from each other? The purpose of a *post hoc* test is to find out exactly where those differences are.

There are several tests that can be employed here, though we will only concern ourselves with two commonly used tests: the Tukey honestly significant difference test, sometimes called the 'Tukey hsd' test, or simply (as in SPSS) the Tukey test, and the Games-Howell test. Others include the Bonferroni and Scheffé test; they are more rigorous than the Tukey test and tend to be used less frequently. The Scheffé test is very similar to the Tukey hsd test, but it is more stringent than the Tukey test in respect of reducing the risk of a Type I error, though this comes with some loss of statistical power (see Chapter 39 on statistical power): one may be less likely to find a difference between groups in the Scheffé test. The Tukey test assumes equality of variances ('homogeneity of variance') in the scores of the groups and equal sub-sample sizes, whilst the Games-Howell test is used if homogeneity of variance is not present or if sub-sample sizes differ.

The Tukey test groups together sub-samples whose means are *not* statistically significantly different from each other and places them in a different group from a group whose means *are* statistically significantly different from the first group. Let us see what this means in our example of the mathematics results of four types of school (Table 41.9).

Table 41.9 takes each type of school and compares it with the other three types, in order to see where there may be statistically significant differences between them. Here the rural primary school is first compared with the rural secondary school (row one of the left-hand column cell named 'Rural primary'), and no statistically significant difference is found between them (Sig. = 0.996, i.e. $\rho > 0.05$). The rural primary school is then compared with the urban primary school and a statistically significant difference is found between them (Sig. = 0.002, i.e. $\rho < 0.05$). The rural primary school is then compared with the urban secondary school, and, again, a statistically significant difference is found between them (Sig. = 0.003, i.e. $\rho < 0.05$). The next cell of the left hand column commences with the

TABLE 41.9 THE TUKEY TEST (SPSS OUTPUT)

Multiple Comparisons

Dependent Variable: Standardised Mathematics scores (percentages)

Tukey HSD

(I) Grouping of school	(J) Grouping of school	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Rural primary	Rural secondary	-.59	2.563	.996	-7.19	6.01
	Urban primary	9.21*	2.541	.002	2.67	15.76
	Urban secondary	8.15*	2.366	.003	2.06	14.24
Rural secondary	Rural primary	.59	2.563	.996	-6.01	7.19
	Urban primary	9.80*	2.531	.001	3.28	16.32
	Urban secondary	8.74*	2.355	.001	2.67	14.81
Urban primary	Rural primary	-9.21*	2.541	.002	-15.76	-2.67
	Rural secondary	-9.80*	2.531	.001	-16.32	-3.28
	Urban secondary	-1.06	2.331	.968	-7.07	4.94
Urban secondary	Rural primary	-8.15*	2.366	.003	-14.24	-2.06
	Rural secondary	-8.74*	2.355	.001	-14.81	-2.67
	Urban primary	1.06	2.331	.968	-4.94	7.07

*. The mean difference is significant at the .05 level.

rural secondary school, and this is compared with the rural primary school, and no statistically significant difference is found (Sig. = 0.996, i.e. $\rho > 0.05$). The rural secondary school is then compared to the urban primary school and a statistically significant difference is found between them (Sig. = 0.001, i.e. $\rho < 0.05$). The rural secondary school is then compared with the urban secondary school, and, again, a statistically significant difference is found between them (Sig. = 0.001, i.e. $\rho < 0.05$). The analysis is continued for the urban primary and the urban secondary school. One can see that the two types of rural school do *not* differ statistically significantly from each other, that the two types of urban school do *not* differ statistically significantly from each other, but that the rural and urban schools *do* differ statistically significantly from each other. We can see where the null hypothesis *is* supported and where it is *not* supported.

In fact the Tukey test in SPSS presents this very clearly, as shown in Table 41.10. (If the Games-Howell test is used then a similar output is provided by SPSS.) In Table 41.10, one group of similar means (i.e. those not statistically significantly different from each other: the urban primary and urban secondary) is placed together (the column labelled '1') and the other group of similar means (i.e. those not statistically significantly different from each other: the rural primary and rural secondary) is placed together (the column labelled '2'). SPSS automatically groups these and places them in ascending order (the group with the lowest means appears in the first column, and the group with the highest means is in the second column). So, one can see clearly that the difference between the school lies *not* in the fact that some are primary and some are secondary, but that some are rural and some are urban, i.e. the differences relate to geographical location rather than age group in the school. The Tukey test helps us to locate exactly where the similarities and differences between groups lie. It places the means into homogeneous sub-groups, so that we can see which means are close together but different from other groups of means.

TABLE 41.10 HOMOGENEOUS GROUPINGS IN THE TUKEY TEST (SPSS OUTPUT)

Standardised Mathematics scores (percentages)

Tukey HSD[a,b]

Grouping of school	N	Subset for alpha = .05	
		1	2
Urban primary	141	50.64	
Urban secondary	194	51.70	
Rural primary	134		59.85
Rural secondary	136		60.44
Sig.		.973	.995

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 147.806.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Analysis of Variance here tells us that there are or are not statistically significant differences between groups; the Tukey test indicates where these differences lie, if they exist. We advise using the two tests together. Of course, as with the t-test, it is sometimes just as important if we do *not* find a difference between groups as if we *do* find a difference. For example, if we were to find that there was no statistically significant difference between four groups (parents, teachers, students and school governors/leaders) on a particular issue, say, the move towards increased science teaching, then this would give us greater grounds for thinking that a proposed innovation – the introduction of increased science teaching – would stand a greater chance of success than if there had been statistically significant differences between the groups. Finding no difference can be as important as finding a difference.

In reporting Analysis of Variance and the Tukey test one could use a form of words thus:

> Analysis of Variance found that there was a statistically significant difference between rural and urban schools ($F_{14} = 8.975$, $\rho < 0.001$). The Tukey test found that the means for rural primary schools and rural secondary schools (59.85 and 60.44 respectively) were not statistically significantly different from each other, and that the means for urban primary schools and urban secondary schools (50.64 and 51.70 respectively) were not statistically significantly different from each other. The homogeneous subsets calculated by the Tukey test reveal two subsets in respect of the variable 'Standardized mathematics scores': (a) urban primary and urban secondary schools; (b) rural primary and rural secondary schools. The two subsets reveal that these two groups were distinctly and statistically significantly different from each other in respect of this variable. The means of the rural schools were statistically significantly higher than the means of the urban schools.

Box 41.3 provides the SPSS command sequence for one-way ANOVA with the Tukey test.

Box 41.4 provides the SPSS command sequence for repeated measures in ANOVA with the Tukey test.

BOX 41.3 SPSS COMMAND SEQUENCE FOR ONE-WAY ANOVA WITH THE TUKEY TEST

To run one-way ANOVA in SPSS, checking for homogeneity of variance, with the Tukey and Games-Howell tests, the command sequence is: 'Analyze' → 'Compare means' → 'One-Way ANOVA' → Send the independent variable to the 'Factor' box → Send the dependent variable to the 'Dependent' box → Click the 'Post Hoc' box → Check the 'Tukey' box and the 'Games-Howell' box → Click 'Continue' → Click the 'Options' box → Check the boxes marked 'Homogeneity of variance test', 'Brown-Forsythe' and 'Welch' → Click 'Continue' (which returns you to the original screen) → Click 'OK'. The Homogeneity of variance test is the test of equality of variance (which informs the researcher whether to proceed with the Tukey or the Games-Howell test), and if the homogeneity of variance is violated (i.e. $\rho > 0.05$) then the Brown-Forsythe test and the Welch tests are more robust than the ANOVA, and SPSS provides the significance levels for these two tests here.

BOX 41.4 SPSS COMMAND SEQUENCE FOR REPEATED MEASURE ANOVA WITH THE TUKEY TEST

For repeated measures in ANOVA using SPSS (i.e. the same groups under three or more conditions), with the Tukey test and a measure of effect size, the SPSS command sequence is: 'Analyze' → 'General Linear Model' → 'Repeated Measures' → In the box 'Within-Subject Factor Name' name a new variable → In the box 'Number of Levels' insert the number of dependent variables that you wish to include→ Click 'Add'→ Click 'Define' → Send over the dependent variables that you wish to include (the number of variables must be the same as the 'Number of Levels') into the box 'Within-Subjects Variables' → Send over the independent variable into the box 'Between-Subjects Factors'→ Click 'Options' → Click 'Estimates of effect size' → Click 'Continue' → Click 'Post Hoc' → Send over the factor from the 'Factor(s)' box to the 'Post Hoc Tests for' box → Click 'Tukey' → Click 'Continue' (which returns you to the original screen) → Click 'OK'. This will give you several boxes in the output. Go to the box called 'Multivariate Tests'; the furthest right-hand column has the partial eta squared. Go to the row that contains the last box, and look at the partial eta squared (e.g. 'Pillai's Trace' and 'Wilk's Lambda'); this gives the partial eta squared.

Statistical significance in one-way ANOVA is an indication of whether the results occur by chance alone. To find how much difference there is (i.e. effect size), the researcher can calculate eta squared, and we refer the reader to Chapter 39, which indicates exactly how this can be calculated. Chapter 39 also reports other tests of effect size, together with how to conduct and interpret them.

Two-way Analysis of Variance

The example of ANOVA above illustrates one-way Analysis of Variance, i.e. the difference between the means of three or more groups on a *single* independent variable. Additionally, ANOVA can take account of more than one independent variable. Two-way Analysis of Variance is used to estimate the effect of two independent variables on a single variable (Cohen and Holliday, 1996, p. 277). Let us take the example of how examination performance in science is affected by both age group and sex. Two-way ANOVA enables the researcher to examine not only the effect of each independent variable but also the interaction effects on each other of these two independent variables, i.e. how sex effects are influenced or modified when combined with age group effects. We may discover, for example, that age group has a differential effect on examination performance according to whether one is male or female, i.e. there is an interaction effect.

For two-way Analysis of Variance the researcher requires two independent categorical (nominal) variables, for example, sex, age group and one continuous dependent variable (e.g. performance on examinations). Two-way ANOVA enables the researcher to calculate three effects. In this example they are:

- differences in examination performance by sex;
- difference in examination performance by age group;
- the interaction of sex and age group on examination, for example, is there a difference in the effects of age group on examination performance for males and females?

We will use SPSS to provide an example of this. SPSS firstly presents descriptive statistics, as shown in Table 41.11. This table simply presents the data, with means and standard deviations. Next SPSS calculates the

TABLE 41.11 MEANS AND STANDARD DEVIATIONS IN A TWO-WAY ANALYSIS OF VARIANCE (SPSS OUTPUT)

Descriptive Statistics

Dependent Variable: SCIENCE

sex	Age group	Mean	Std. Deviation	N
male	15-20	71.92	24.353	125
	21-25	63.33	31.459	111
	26-45	70.95	28.793	21
	46 and above	64.69	28.752	128
	Total	66.99	28.390	385
female	15-20	70.33	25.768	182
	21-25	68.82	25.396	221
	26-45	69.59	28.059	49
	46 and above	61.66	28.464	163
	Total	67.43	26.731	615
Total	15-20	70.98	25.173	307
	21-25	66.99	27.646	332
	26-45	70.00	28.079	70
	46 and above	62.99	28.581	291
	Total	67.26	27.369	1000

Levene test for equality of error variances, degrees of freedom and significance levels (Table 41.12).

The Levene test in Table 41.12 enables the researcher to know whether there is equality across the variances. The researcher needs to see if the significance level is greater than 0.05, looking for a significance level *greater than* 0.05, i.e. *not* statistically significant, which supports the null hypothesis which states that there is no statistically significant difference between the variances across the groups (i.e. to support the assumptions of ANOVA). In our example this is the case as the significance level is 0.156. The researcher is safe, therefore, to proceed with the analysis. SPSS provides here with important information, as shown in Table 41.13.

TABLE 41.12 THE LEVENE TEST OF EQUALITY OF VARIANCES IN A TWO-WAY ANALYSIS OF VARIANCE (SPSS OUTPUT)

Levene's Test of Equality of Error Variances [a]

Dependent Variable: SCIENCE

F	df1	df2	Sig.
3.463	7	992	.001

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+SEX+AGE GROUP +SEX * AGE GROUP

In this table, there are three sets of independent variables listed (SEX, AGE GROUP, SEX*AGE GROUP). The column headed 'Sig.' shows that the significance levels for the three sets are, respectively: 0.956, 0.004 and 0.244. Sex does not have a statistically significant effect on science examination performance ($\rho = 0.956$). Age group does have a statistically significant effect on the performance in the science examination ($\rho = 0.004$). The interaction effect of sex and age group does not have a statistically significant effect on performance ($\rho = 0.244$). SPSS also computes the effect size (Partial Eta squared). For the important variable AGE GROUP this is given as 0.014, which shows that the effect size is very small indeed, suggesting that even though statistical significance has been found, the actual difference in the mean values is very small. This latter is a neat illustration of

TABLE 41.13 BETWEEN-SUBJECT EFFECTS IN TWO-WAY ANALYSIS OF VARIANCE (SPSS OUTPUT)

Tests of Between-Subjects Effects

Dependent Variable: SCIENCE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
Corrected Model	13199.146[b]	7	1885.592	2.545	.013	.018	17.812	.888
Intercept	2687996.888	1	2687996.9	3627.42	.000	.785	3627.421	1.000
SEX	2.218	1	2.218	.003	.956	.000	.003	.050
AGE GROUP	10124.306	3	3374.769	4.554	.004	.014	13.663	.887
SEX * AGE GROUP	3089.630	3	1029.877	1.390	.244	.004	4.169	.371
Error	735093.254	992	741.021					
Total	5272200.000	1000						
Corrected Total	748292.400	999						

a. Computed using alpha = .05

b. R Squared = .018 (Adjusted R Squared = .011)

the point made in Chapter 39, that relying on statistical significance alone is dangerous, and that effect sizes are often more useful and, indeed, tell us *how much* of a difference there is, which is something that significance testing alone cannot do.

As with one-way ANOVA, the Tukey and/or the Games-Howell tests can be applied here to present the homogeneous groupings of the sub-sample means. SPSS can also present a graphic plot of the two sets of scores, which gives the researcher a ready understanding of the effects of the males and females across the four age groups in their science examination (Figure 41.1).

In reporting the results of the two-way Analysis of Variance one can use the following form of words:

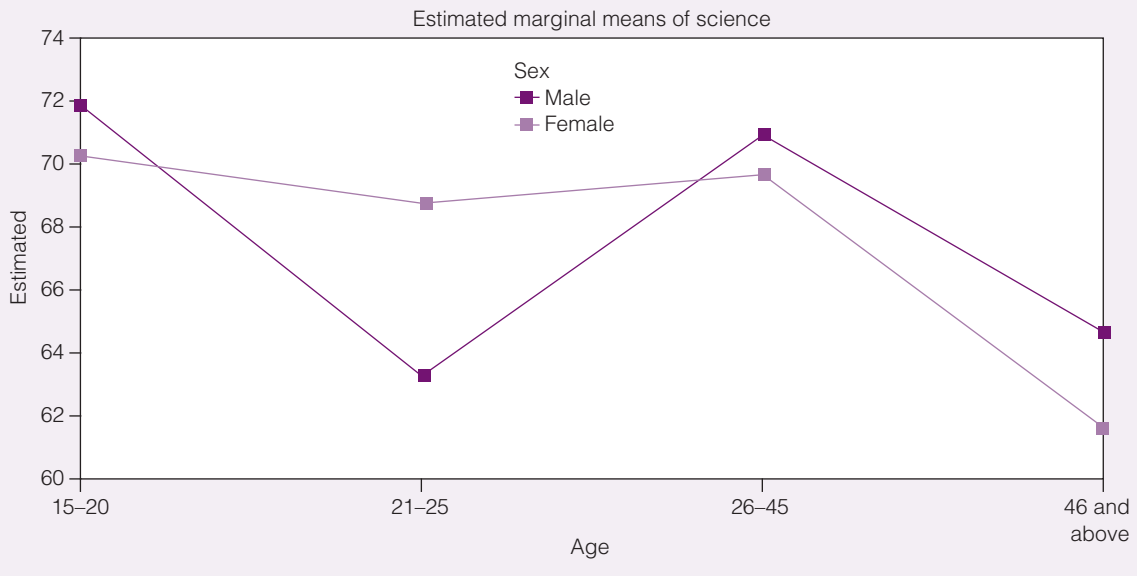


FIGURE 41.1 Graphic plots of two sets of scores on a dependent variable

A two-way between-groups Analysis of Variance was conducted to discover the impact of sex and age group on performance in a science examination. Subjects were divided into four groups by age: Group 1: 15–20 years; Group 2: 21–25 years; Group 3: 26–45 years; and Group 4: 46 years and above. There was a statistically significant main effect for age group ($F=4.554$, $\rho=0.004$); however, the effect size was small (partial eta squared=0.014). The main effect for sex ($F=0.003$, $\rho=0.956$) and the interaction effect ($F=1029.877$, $\rho=0.244$) were not statistically significant.

Box 41.5 provides the SPSS command sequence for two-way ANOVA.

Multiple Analysis of Variance

Multiple Analysis of Variance (MANOVA) is designed to see the effects of one categorical independent variable on two or more continuous variables (e.g. 'do males score more highly than females in terms of how hard they work and their IQ'). We mention it here by way of introduction, but we refer readers to more advanced texts (e.g. Tabachnick and Fidell, 2013; Pallant, 2016), for a fuller analysis of how to conduct this.

Safety checks for using MANOVA

MANOVA is very sensitive to the assumptions that are made about the data, and researchers should conduct 'safety checks' to ensure that the data are suitable for this statistic to be calculated:

- continuous parametric data for dependent variables;
- independent variables are categorical, with two or more values;
- groups are independent of each other;
- random sampling;
- adequate sample size (more cases in each cell than the number of dependent variables being studied, e.g. a minimum of twenty cases in each cell);
- normal distribution of the data (though large samples often overcome this);
- no outliers;
- a linear relationship between each pair of dependent variables;
- no multicollinearity (dependent variables are independent of each other but moderately correlated);
- homogeneity (equality) of variances (though the Levene test can identify problems here, and SPSS can offer the Brown-Forsythe and Welch tests to overcome the problem here, discussed below).

Box 41.6 provides the SPSS command sequence for MANOVA.

BOX 41.5 SPSS COMMAND SEQUENCE FOR TWO-WAY ANOVA

To run two-way ANOVA in SPSS the command sequence is: 'Analyze' → Click 'General Linear Model' → Click 'Univariate' → Send the dependent variable to the box 'Dependent Variable' → Send the independent variables to the box 'Fixed Factors' → Click the 'Options' box; in the 'Display' area, check the boxes 'Descriptive Statistics' and 'Estimates of effect size' → Click 'Continue' → Click the 'Post Hoc' box, and send over from the 'Factors' box to the 'Post Hoc tests for' box those factors that you wish to investigate in the post hoc tests → Click the post hoc test that you wish to use (e.g. Tukey') → Click 'Continue' → Click the 'Plots' box → Move to the 'Horizontal' box the factor that has the most groups → Move to the 'Separate lines' box the factor the other independent variable → Click 'Add' → Click 'Continue' → Click 'OK'.

BOX 41.6 SPSS COMMAND SEQUENCE FOR MANOVA

To run MANOVA, the SPSS command sequence is: 'Analyze' → Click 'General Linear Model' → Click on 'Multivariate' → Send to the box 'Dependent Variables' the dependent variables that you wish to include → Send to the 'Fixed Factors' box the independent variable that you wish to use → Click 'Model' and ensure that the 'Full factorial' (in the 'Specify Model' box) and the 'Type III' (in the 'Sum of Squares') boxes are selected → Click 'Continue' → Click the 'Options' box and send to the 'Display Means for' box the independent variable that you wish to include → In the 'Display' section, click 'Descriptive Statistics', 'Estimates of effect size' and 'Homogeneity tests' → Click 'Continue' → If you want a post hoc test click 'Post Hoc' → Send over the independent variable from the 'Factor(s)' box to the 'Post Hoc tests for' box (if you want a post hoc test and if your independent variable has three or more values) → Click 'Tukey' and 'Games-Howell' → Click 'Continue' → Click 'OK'.

For further guidance on running SPSS for these matters and interpreting the SPSS output, we refer readers to Tabachnick and Fidell (2013) and Pallant (2016).

41.4 The chi-square test

Difference testing is an important feature in educational research. We can conduct a chi-square test (χ^2) (pronounced 'kigh', as in 'high') to investigate difference. The chi-square test is a test of difference that can be conducted for a univariate analysis (one categorical variable), and between two categorical variables. The chi-square test measures the difference between a statistically generated expected result and an actual (observed) result to see if there is a statistically significant difference between them, i.e. to see if the frequencies observed are statistically significantly different or by chance alone; it is a measure of 'goodness of fit' between an expected and an actual, observed result or set of results. The expected result is based on a statistical process discussed below. Here is not the place to go into the mathematics of the test, not least because computer packages automatically calculate the results, though the formula for calculating chi-square is:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where
O = observed frequencies
E = expected frequencies
$\sum$ = the sum of

For *univariate data* (one variable) let us take the example of 120 students who were asked which of four teachers they preferred. We start with the null hypothesis that states that there is no statistically significant difference in their preferences for four teachers, i.e. that the 120 scores are spread evenly across the four teachers (30 votes for each teacher), thus:

Frequencies	*Teacher A*	*Teacher B*	*Teacher C*	*Teacher D*
Observed	20	70	10	20
Expected	30	30	30	30
Residual (difference between observed and expected frequencies)	–10	40	–20	–10

The results indicate that the assumed equal distribution of preferences (30 for each teacher) in reality is not evenly distributed; there are 20 for teacher A, 70 for teacher B, 10 for teacher C and 20 for teacher D. Is this by chance or are these differences statistically significant? Using the formula above we compute the chi-square value thus:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$
$$= \sum \frac{(-10)^2}{30} + \frac{(40)^2}{30} + \frac{(-20)^2}{30} + \frac{(-10)^2}{30}$$
$$= \frac{100+1600+400+100}{30} = \frac{2200}{30} = 73.3$$

The chi-square value here is 73.3, with three degrees of freedom (explained below). In a table of critical values for chi-square distributions (in the appendices of most statistics books and freely available on the Internet), we look up the level of statistical significance for three degrees of freedom:

Degrees of freedom	*Level of significance*	
	0.05	*0.01*
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81

We find that, at 73.3, the chi-square value is considerably larger than the 11.34 given in the table, i.e. it has a probability level which is more than by chance, being stronger than 0.01, i.e. there is a statistically significant difference between the observed and expected frequencies, i.e. not all teachers are equally preferred (more people preferred Teacher B, and this was statistically significant, i.e. not by chance).

Box 41.7 provides the SPSS command sequence for univariate chi-square.

For *bivariate data* (two variables) the chi-square test is a test of independence, to see whether there is a relationship or association between two categorical variables. Let us say that we have a crosstabulation of males and females and their liking for maths (like/dislike). We start with the null hypothesis that states that there is *no* statistically significant difference between the males and females (variable 1) in their (dis)liking for mathematics (variable 2), and we seek to discover if the null hypothesis is supported.

BOX 41.7 SPSS COMMAND SEQUENCE FOR UNIVARIATE CHI-SQUARE

In SPSS the command sequence for univariate chi-square is: 'Analyze' → 'Nonparametric tests' → 'Legacy dialogs' → 'Chi-square' → Send over the variable of interest to the 'Test variable' box → Click 'OK'. Note that the 'expected values' default setting is 'All categories equal'. If the researcher expects that not all the categories will be equal then she/he can set the expected distribution proportions in the 'expected values' window by setting the values as a decimal fraction, for example, 0.6, 0.3, 0.1 for three categories (which must not exceed 1.0 in total).

We set the level of significance (α) that we wish to use for supporting or not supporting the null hypothesis; for example we could say 'Let $\alpha=0.05$'. Having found out the true voting we set out a 2×2 crosstabulation thus, with the observed frequencies in the cells:

	Male	*Female*	*Total*
Like mathematics	60	25	85
Dislike mathematics	35	75	110
Total	95	100	195

These are the observed frequencies. To find out the expected frequencies for each cell we use the formula:

$$\text{Expected value of a cell} = \frac{\text{row total} \times \text{column total}}{\text{Overall total}}$$

Using the row totals, the column totals and the overall total, we can calculate the expected frequencies for each thus (figures rounded):

	Male	*Female*	*Total*
Like mathematics	(85 × 95)/ 195=41.4	(85 × 100)/ 195=43.6	85
Dislike mathematics	(110 × 95)/ 195=53.6	(110 × 100)/ 195=56.4	110
Total	95	100	195

The chi-square value, using the formula above is:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \sum \frac{(60-41.4=18.6)^2}{41.4} + \frac{(25-43.6=18.6)^2}{53.6} + \frac{(35-53.6=18.6)^2}{43.6} + \frac{(75-56.4=18.6)^2}{56.4}$$

$$= \frac{345.96}{41.4} + \frac{345.96}{53.6} + \frac{345.96}{43.6} + \frac{345.96}{56.4}$$

$$= 8.36 + 6.45 + 7.93 + 6.13$$

$$= 28.87$$

When we look up the chi-square value of 28.87 in the table of critical values of the chi-square distribution earlier, with two degrees of freedom, we observe that the figure of 28.87 is larger than the figure of 9.21 given in that table and required for statistical significance at the 0.01 level. Hence we conclude that the distribution of likes and dislikes for mathematics by males and females is not simply by chance but that there is a statistically significant difference between the voting of males and females here. Hence the null hypothesis is not supported and the alternative hypothesis, that there is a statistically significant difference between the voting of the two groups, is supported.

We do not need to perform these calculations by hand. Computer software such as SPSS will do all of the calculations with a few keystrokes.

We recall that the conventionally accepted minimum level of statistical significance is usually 0.05, and we used this level in the example; the significance level of our data here is smaller than either the 0.05 and 0.01 levels, i.e. it is highly statistically significant.

One can report the results of the chi-square test thus:

> When the chi-square statistic was calculated for the distribution of males and females on their liking for mathematics, a statistically significant difference was found between the males and the females ($\chi^2=28.87$, $df=2$, $\rho=0.000$).

We use Yates's correction (a continuity correction) to compensate for the over-estimate of the chi-square in a 2×2 table, and this can be activated by simple keystrokes in SPSS or other software.

The chi-square statistic is normally used with nominal (categorical) data, and our example illustrated this. We provide a further example of the chi-square statistic, with data that are set into a contingency table, this time in a 2×3 contingency table, i.e. two horizontal rows and three columns (contingency tables may contain more than this number of variables). The example this time presents data concerning sixty students' entry into science, arts and humanities, in a college, and whether the students are male or female.

The lower of the two figures in each cell is the number of actual students who have opted for the particular subjects (sciences, arts, humanities). The upper of the two figures in each cell is what might be expected purely by chance to be the number of students opting for each of the particular subjects. The figure is arrived at by statistical computation, hence the decimal fractions for the figures. What is of interest to the researcher is whether the actual distribution of subject choice by males and females differs significantly from that which could occur by chance variation in the population of college entrants (Table 41.14).

The researcher begins with the null hypothesis that there is no statistically significant difference between the actual results noted and what might be expected to occur by chance in the wider population. When the chi-square statistic is calculated, if the observed, actual distribution differs from that which might be expected to occur by chance alone, then the researcher has to determine whether that difference is statistically significant, i.e. not to support the null hypothesis.

In our example of sixty students' choices, the chi-square formula yields a final chi-square value of 14.64. This we refer to the tables of the critical values of the chi-square distribution (an extract from which is set out for the first example above) to determine whether the derived chi-square value indicates a statistically significant difference from that occurring by chance.

TABLE 41.14 A 2 × 3 CONTINGENCY TABLE FOR CHI-SQUARE

	Science subjects	*Arts subjects*	*Humanities subjects*	
Males	7.6 14	8 4	8.4 6	24
Females	11.4 5	12 16	12.6 15	36
	19	20	21	60

The researcher sees that the 'degrees of freedom' (a mathematical construct that is related to the number of restrictions that have been placed on the data) has to be identified. In many cases, to establish the degrees of freedom, one simply takes 1 away from the total number of rows of the contingency table and 1 away from the total number of columns and adds them; in this case it is (2–1)+(3–1)=3 degrees of freedom. Degrees of freedom are discussed in the next section. (Other formulae for ascertaining degrees of freedom hold that the number is the total number of cells minus 1.) The researcher looks along the table from the entry for the three degrees of freedom and notes that the derived chi-square value calculated (14.64) is statistically significant at the 0.01 level, i.e. is higher than the required 11.34, indicating that the results obtained – the distributions of the actual data – could not have occurred simply by chance. The null hypothesis is not supported at the 0.01 level of statistical significance. Interpreting the specific numbers of the contingency table (Table 41.14) in educational rather than statistical terms, noting (a) the low incidence of females in the science subjects and the high incidence of females in the arts and humanities subjects, and (b) the high incidence of males in the science subjects and low incidence of males in the arts and humanities subjects, the researcher would say that this distribution is statistically significant – suggesting, perhaps, that the college needs to consider action possibly to encourage females into science subjects and males into arts and humanities.

The chi-square test is one of the most widely used tests, and is applicable to nominal data in particular. More powerful tests are available for ordinal, interval and ratio data, and we discuss these separately. However, one has to be cautious of the limitations of the chi-square test. Look at the example shown in Table 41.15.

TABLE 41.15 A 2 × 5 CONTINGENCY TABLE FOR CHI-SQUARE

	Music	*Physics*	*Maths*	*German*	*Spanish*	
Males	7 14.0%	11 22.0%	25 50%	4 8.0%	3 6%	50 100%
Females	17 12.1%	38 27.0%	73 52%	12 8.5%	1 0.7%	141 100%
Total	24 12.6%	49 25.7%	98 51%	16 8.4%	4 2.1%	191 100%

If one were to perform the chi-square test on Table 41.15 then one would have to be very cautious. The chi-square statistic assumes that no more than 20 per cent of the total number of cells contain fewer than five cases. In the example here we have one cell with four cases, another with three, and another with only one case, i.e. three cells out of the ten (two rows – males and females – with five cells in each for each of the rating categories). This means that 30 per cent of the cells contain fewer than five cases; even though a computer will calculate a chi-square statistic, it means that the result is unreliable. This highlights the point made in Chapter 12 about sampling, namely. that the subsample size has to be large. For example, if each category here were to contain five cases then it would mean that the minimum sample size would be fifty (10×5), assuming that the data are evenly spread. In the example here, even though the sample size is much larger (191) it still does not guarantee that the 20 per cent rule will be observed, as the data are unevenly spread. When calculating the chi-square statistic, the researcher can use the Fisher's Exact Probability Test if more than 25 per cent of the cells have fewer than five cases, and this is automatically calculated and printed as part of the normal output in the chi-square calculation in SPSS.

Because of the need to ensure that at least 80 per cent of the cells of a chi-square contingency table contain more than five cases if confidence is to be placed in the results, it may not be feasible to calculate the chi-square statistic if only a small sample is being used. Hence the researcher would tend to use this statistic for larger-scale survey data. Other tests could be used if the problem of low cell frequencies obtains, for example, the binomial test and, more widely used, the Fisher Exact Probability Test (Cohen and Holliday, 1996, pp. 218–20). The required minimum number of cases in each cell renders the chi-square statistic problematical, and, apart from with nominal data, there are alternative statistics that can be calculated and which overcome this problem (e.g. the Mann-Whitney, Wilcoxon, Kruskal-Wallis and Friedman tests for non-parametric – ordinal – data, and the t-test and Analysis of Variance test for parametric – interval and ratio – data).

With statistical significance being increasingly questioned and being replaced with measures of effect size (see Chapter 39), calculations of effect size for categorical tables (crosstabulations) use two main statistics (see Chapter 39):

- the *phi* coefficient for 2×2 tables (in which Cohen's d indicates small effect for 0.10, a medium effect for 0.30 and a large effect for 0.50).
- Cramer's V for contingency tables larger than 2×2, which takes account of degrees of freedom.

Two significance tests for very small samples are given in the accompanying website.

Box 41.8 provides the SPSS command sequence for bivariate chi-square with crosstabulations for simple frequencies, i.e. in which each case has its own row in the SPSS file. However, sometimes the data that researchers collect do not come in a case-by-case form but are already aggregated, i.e. with totals rather than individual cases, and Box 41.9 indicates how to work with this in SPSS.

41.5 Degrees of freedom

The chi-square statistic introduces the term *degrees of freedom*. Gorard (2001b, p. 233) suggests that 'the degrees of freedom is the number of scores we need to know before we can calculate the rest'. Cohen and Holliday (1996) explain the term clearly:

> Suppose we have to select any five numbers. We have complete freedom of choice as to what the numbers are. So, we have five degrees of freedom. Suppose however we are then told that the five

BOX 41.8 SPSS COMMAND SEQUENCE FOR BIVARIATE CHI-SQUARE WITH CROSSTABULATIONS

The SPSS command sequence for bivariate chi-square works with the 'Crosstabs' command; it is: 'Analyze' → 'Descriptive Statistics' → 'Crosstabs' → Send over the row variable to the 'Rows' box and send over the column variables to the 'Columns' box → Click the 'Statistics' box, which opens a new window → Check the 'Chi-square' and the 'Phi and Cramer's V' boxes → Click 'Continue' → Click the 'Cells' box, which opens a new window → In the 'Percentages' area of the new window check the 'Total' box → Click 'Continue' → Click 'OK'. Note that SPSS automatically applies Yates's correction ('Continuity Correction') for 2×2 tables in chi-square and SPSS automatically uses the Fisher Exact Probability test if more than 25 per cent of the cells have fewer than five cases.

BOX 41.9 SPSS COMMAND SEQUENCE FOR BIVARIATE CHI-SQUARE WITH AGGREGATED DATA

Sometimes the researcher only has already-aggregated data. For example, imagine that the researcher is given only a table of aggregated totals of subject choices by 200 students (100 males and 100 females) at age 15, thus:

	Subject choices			
	Physical sciences	*Arts and humanities*	*Social sciences*	*Business studies*
Males	27	12	18	43
Females	10	29	28	33

Here each row contains aggregated cases rather than single cases. SPSS can handle this but the data have to be entered differently in SPSS and a preliminary command sequence for weighting cases has to take place before the command sequence in Box 41.8 can be followed. Here there are eight cells of numbers, with one variable for 'Gender' (males and females) and one variable for 'subject choices'. SPSS creates a variable for 'Gender', with values in the SPSS file as: 1='Males', 2='Females'. SPSS creates a variable for 'Subject choices', with values in the SPSS files as: 1='Physical Sciences', 2='Arts and Humanities', 3='Social Sciences' and 4='Business Studies'. SPSS creates a new variable of 'Frequencies' to contain the numbers in each cell; this is a scale variable in SPSS. Having created the SPSS file the 'Data View' window in SPSS takes each of the pieces of data for the eight cells in the first table above and reforms it for SPSS, to appear thus:

Gender	*Subject choices*	*Frequencies*
1	1	27
1	2	12
1	3	18
1	4	43
2	1	10
2	2	29
2	3	28
2	4	33

Data row one is for males choosing Physical Sciences, with 27 students.
Data row two is for males choosing Arts and Humanities, with 12 students.
Data row three is for males choosing Social Sciences, with 18 students.
Data row four is for males choosing Business Studies, with 43 students.
Data row one is for females choosing Physical Sciences, with 10 students.
Data row two is for females choosing Arts and Humanities, with 29 students.
Data row three is for females choosing Social Sciences, with 28 students.
Data row four is for females choosing Business Studies, with 33 students.

To run a chi-square in SPSS here, the researcher constructs this data file and then runs the 'Weight Cases' command sequence: 'Data' → 'Weight Cases' → Click the radio button 'Weight cases by' and send the variable 'Frequencies' to the box marked 'Frequency variable' → Click 'OK'. Having done this, the SPSS command sequence in Box 41.8 can be followed.

numbers must have a total value of 25. We will have complete freedom of choice to select four numbers but the fifth will be dependent on the other four. Let's say that the first four numbers we select are 7, 8, 9, and 10, which total 34, then if the total value of the five numbers is to be 25, the fifth number must be –9.

$$7+8+9+10-9=25$$

A restriction has been placed on one of the observations; only four are free to vary; the fifth has lost its freedom. In our example then d.f. = 4, that is $N-1=5-1=4$.

Suppose now that we are told to select any five numbers, the first two of which have to total 9, and the total value of all five has to be 25. Our restriction is apparent when we wish the total of the first two numbers to be 9. Another restriction is apparent in the requirement that all five numbers must total 25. In other words we have lost two degrees of freedom in our example. It leaves us with d.f. = 3, that is, $N-2=5-2=3$.'

(Cohen and Holliday, 1996, p. 113)

For a crosstabulation (a contingency table), degrees of freedom refer to the freedom with which the researcher is able to assign values to the cells, given fixed marginal totals, usually given as (number of rows – 1)+ (number of columns – 1). There are many variants of this, and readers will need to consult more detailed texts to explore this issue. We do not dwell on degrees of freedom here, as it is automatically calculated and addressed in calculations by most statistical software packages such as SPSS.

41.6 The Mann-Whitney and Wilcoxon tests

The non-parametric equivalents of the t-test are the Mann-Whitney U test for two independent samples and the Wilcoxon test for two related samples, both used with one categorical variable and a minimum of one ordinal variable. These enable us to see, for example, whether there are statistically significant differences between males and females on a rating scale.

The Mann-Whitney U test is based on ranks – how many times a score from one group is ranked higher than a score from another group (Bryman and Cramer, 1990, p. 129), thereby overcoming the problem of low cell frequencies in the chi-square statistic. Let us take an example. Imagine that we have conducted a course evaluation, using five-point rating scales ('very little'; 'a little'; 'a moderate amount'; 'quite a lot'; 'a very great deal'), and we wish to find if there is a statistically significant difference between the voting of males and females on the variable 'The course gave you opportunities to learn at your own pace', i.e. whether any differences between males and females are by chance alone. We commence with the null hypothesis ('there is no statistically significant difference between the two rankings') and then we set the level of significance (α) for supporting or not supporting the null hypothesis; for example we could say 'Let $\alpha = 0.05$'. A crosstabulation is shown in Table 41.16.

Are the differences between the two groups statistically significant? Using SPSS, the Mann-Whitney statistic is given in Tables 41.17 and 41.18.

Mann-Whitney using ranks (as in Table 41.17) yields a U-value of 2,732.500 (Table 41.18) from the formula it uses for the calculation (SPSS does this automatically). The important information in Table 41.18 is the 'Asymp. Sig. (2-tailed)', i.e. the statistical significance level of

TABLE 41.16 A CROSSTABULATION FOR A MANN-WHITNEY U TEST (SPSS OUTPUT)

sex * the course gave you opportunities to learn at your own pace Crosstabulation

		the course gave you opportunities to learn at your own pace					
		strongly disagree	disagree	neither agree nor disagree	agree	strongly agree	Total
male	Count	1	2	16	21	9	49
	% within sex	2.0%	4.1%	32.7%	42.9%	18.4%	100.0%
female	Count	4	11	61	57	8	141
	% within sex	2.8%	7.8%	43.3%	40.4%	5.7%	100.0%
Total	Count	5	13	77	78	17	190
	% within sex	2.6%	6.8%	40.5%	41.1%	8.9%	100.0%

TABLE 41.17 SPSS OUTPUT ON RANKINGS FOR THE MANN-WHITNEY U TEST (SPSS OUTPUT)

Ranks

	sex	N	Mean Rank	Sum of Ranks
the course gave you opportunities to learn at your own pace	male	49	110.23	5401.50
	female	141	90.38	12743.50
	Total	190		

TABLE 41.18 THE MANN-WHITNEY U VALUE AND SIGNIFICANCE LEVEL (SPSS OUTPUT)

Test Statistics [a]

	the course gave you opportunities to learn at your own pace
Mann-Whitney U	2732.500
Wilcoxon W	12743.500
Z	-2.343
Asymp. Sig. (2-tailed)	.019

a. Grouping Variable: sex

any difference found between the two groups (males and females). Here the significance level ($\rho = 0.019$, i.e. $\rho < 0.05$) indicates that the voting by males and females is statistically significantly different and that the null hypothesis is not supported, in other words the differences were not simply by chance. In the t-test and the Tukey test, researchers could immediately find exactly where differences might lie between the groups (by looking at the means and the homogeneous sub-groups respectively). Unfortunately the Mann-Whitney U test does not enable the researcher to identify clearly where the differences lie between the two groups, so the researcher would need to go back to the crosstabulation to identify where differences lie. In the example above, it appears that the males feel more strongly than the females that the course in question has afforded them the opportunity to learn at their own pace.

In reporting the Mann-Whitney U test one could use a form of words such as the following:

> When the Mann-Whitney statistic was calculated to determine whether there was any statistically significant difference in the voting of the two groups ($U = 2{,}732.500$, $\rho = 0.019$), a statistically significant difference was found between the males and females. A crosstabulation found that males felt more strongly than the females that the course in question had afforded them the opportunity to learn at their own pace.

Box 41.10 provides the SPSS command sequence for the Mann-Whitney statistic.

For two *related* samples (e.g. the same group voting for more than one item, or the same grouping voting at

BOX 41.10 SPSS COMMAND SEQUENCE FOR THE MANN-WHITNEY STATISTIC

The SPSS command sequence for the Mann-Whitney statistic is: 'Analyze' → 'Nonparametric statistics' → 'Legacy Dialogs' → '2 Independent Samples' → Send the dependent variable to box 'Test variable list' and the independent variable to the 'Grouping variable' box → Click 'Define groups' (which is activated when the 'Grouping variable' box contains the independent variable) and then type the number that you assigned to each of the two groups in the SPSS file (e.g. males '1' and females '2') → Click 'Continue' (which returns you to the original screen) → Click 'OK'.

two points in time), the Wilcoxon test is applied, and the data are presented and analysed in the same way as the Mann-Whitney U test. For example, in Tables 41.19 and 41.20 there are two variables ('The course was just right' and 'The lecturer was well prepared'), voted on by the same group. The frequencies are given. Is there a statistically significant difference in the voting for these two variables?

As it is the *single, same group* voting on two variables, the sample is not independent, hence the Wilcoxon test is used. Using SPSS output, the data analysis shows that the voting of the group on the two variables is statistically significantly different (see Tables 41.21 and 41.22).

The reporting of the results of the Wilcoxon test can be as follows:

> When the Wilcoxon statistic was calculated to determine whether there was any statistically significant difference in the voting of the group on the two variables ($\rho = 0.000$), a statistically significant difference was found. The group was more

TABLE 41.19 FREQUENCIES AND PERCENTAGES OF VARIABLE ONE IN A WILCOXON TEST (SPSS OUTPUT)

the course was just right

	Valid					
	strongly disagree	disagree	neither agree nor disagree	agree	strongly agree	Total
Frequency	2	14	65	76	34	191
Valid Percent	1.0	7.3	34.0	39.8	17.8	100.0

TABLE 41.20 FREQUENCIES AND PERCENTAGES OF VARIABLE TWO IN A WILCOXON TEST (SPSS OUTPUT)

the lecturer was well prepared

	Valid						
	strongly disagree	disagree	neither agree nor disagree	agree	strongly agree	Total	Total
Frequency	3	5	25	85	72	190	191
Valid Percent	1.6	2.6	13.2	44.7	37.9	100.0	

TABLE 41.21 RANKS AND SUMS OF RANKS IN A WILCOXON TEST (SPSS OUTPUT)

Ranks

		N	Mean Rank	Sum of Ranks
the lecturer was well prepared - the course was just right	Negative Ranks	20 [a]	50.30	1006.00
	Positive Ranks	89 [b]	56.06	4989.00
	Ties	81 [c]		
	Total	190		

a. the lecturer was well prepared < the course was just right

b. the lecturer was well prepared > the course was just right

c. the course was just right = the lecturer was well prepared

TABLE 41.22 SIGNIFICANCE LEVEL IN A WILCOXON TEST (SPSS OUTPUT)

Test Statistics[b]

	the lecturer was well prepared - the course was just right
Z	-6.383[a]
Asymp. Sig. (2-tailed)	.000

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

positive about the variable 'The lecturer was well prepared' than for the variable 'The course was just right'.

Box 41.11 provides the SPSS command sequence for running the Wilcoxon test.

For both the Mann-Whitney and Wilcoxon tests, *not* finding a statistically significant difference between groups can be just as important as finding a statistically significant difference between them, as the former suggests that nominal characteristics of the sample make no statistically significant difference to the voting, i.e. the voting is consistent, regardless of particular features of the sample. Both of these tests yield statistical significance alone, not effect size; they are simply measures of chance.

41.7 The Kruskal-Wallis and Friedman tests

The non-parametric equivalents of Analysis of Variance are the Kruskal-Wallis test for three or more independent samples and the Friedman test for three or more related samples, both of them for use with one categorical variable and one ordinal variable. These enable us to see, for example, whether there are differences between three or more groups (e.g. classes, schools, groups of teachers) on a rating scale.

These tests operate in a very similar way to the Mann-Whitney test, being based on rankings. Let us take an example. Teachers in four different groups, according to the number of years that they have been teaching, have been asked to evaluate one aspect of a particular course that they have attended ('The teaching and learning tasks and activities consolidate learning through application'). One of the results is the crosstabulation shown in Table 41.23. Are the groups of teachers statistically significantly different from each other in respect of their voting? We commence with the null hypothesis ('there is no statistically significant difference between the four groups') and then we set the level of significance (α) to use for supporting or not supporting the null hypotheses; for example we could say 'Let $\alpha = 0.05$'.

The Kruskal-Wallis test calculates and presents the results in SPSS as shown in Tables 41.24 and 41.25.

BOX 41.11 SPSS COMMAND SEQUENCE FOR THE WILCOXON TEST

To run the Wilcoxon test for related (paired) samples in SPSS it is important, first, for the researcher to define the single group to be observed under the two conditions. The single group might be, for example, only the males from a total sample of males and females. Here SPSS requires you to use the Select Cases function (the command sequence in SPSS is: 'Data' → 'Select Cases' → Then decide which radio button you wish to activate ('If the condition is satisfied'; 'random sample of cases'; 'Based on time or case range'; 'Use filter variable'), and each of these open another box for further selection and instructions (cf. Pallant, 2016). Once you have selected the cases (NB if you do not use this function then the entire sample is used) the SPSS command sequence for the Wilcoxon test is: 'Analyze' → 'Nonparametric statistics' → 'Legacy Dialogs' → '2 Related Samples' → Click the first variable in which you are interested and send it to the 'Variable 1' box, and then click the second variable in which you are interested and send it to the 'Variable 2' box → Ensure that the box marked 'Wilcoxon' has been checked → Click 'OK'.

TABLE 41.23 CROSSTABULATION FOR THE KRUSKAL-WALLIS TEST (SPSS OUTPUT)

number of years teaching * the teaching and learning tasks and activities consolidate learning through application Crosstabulation

			the teaching and learning tasks and activities consolidate learning through application				
			disagree	neither agree nor disagree	agree	strongly agree	Total
number of years teaching	<16	Count		2	3		5
		% within number of years teaching		40.0%	60.0%		100.0%
	16-18	Count		29	52	14	95
		% within number of years teaching		30.5%	54.7%	14.7%	100.0%
	19-21	Count	6	40	34	7	87
		% within number of years teaching	6.9%	46.0%	39.1%	8.0%	100.0%
	>21	Count			2		2
		% within number of years teaching			100%		100.0%
Total		Count	6	71	91	21	189
		% within number of years teaching	3.2%	37.6%	48.1%	11.1%	100.0%

TABLE 41.24 RANKINGS FOR THE KRUSKAL-WALLIS TEST (SPSS OUTPUT)

Ranks

	number of years teaching	N	Mean Rank
the teaching and learning tasks and activities consolidate learning through application	<16	5	90.60
	16-18	95	106.53
	19-21	87	82.02
	>21	2	123.00
	Total	189	

The important figure to note here is the 0.009 ('Asymp.Sig.) in Table 41.25: the significance level. Because this is less than 0.05 we can conclude that the null hypothesis ('there is no statistically significant difference between the voting by the different groups of years in teaching') is not supported, i.e. that the difference in the voting according to the number of years in teaching by the voters is not simply by chance. As with the Mann-Whitney test, the Kruskal-Wallis test tells us only that there *is* or *is not* a statistically significant difference, not *where* the difference lies. To find out where the difference lies, one has to return to the crosstabulation (Table 41.23) and examine it. In the example here it appears that those teachers in the group which had been teaching for 16–18 years are the most positive about the aspect of the course in question.

In reporting the Kruskal-Wallis test one could use a form of words such as the following:

TABLE 41.25 SIGNIFICANCE LEVELS IN A KRUSKAL-WALLIS TEST (SPSS OUTPUT)

Test Statistics [a,b]

	the teaching and learning tasks and activities consolidate learning through application
Chi-Square	11.595
df	3
Asymp. Sig.	.009

a. Kruskal Wallis Test

b. Grouping Variable: number of years teaching

When the Kruskal-Wallis statistic was calculated to determine whether there was any statistically significant difference in the voting of the four groups ($\chi^2 = 11.595$, $\rho = 0.009$), a statistically significant difference was found between the groups which had different years of teaching experience. A crosstabulation found that those teachers in the group who had been teaching for 16–18 years were the most positive about the variable 'The teaching and learning tasks and activities consolidate learning through application'.

The k-sample slippage test from Conover (1971), as an alternative to the Kruskal-Wallis test, is set out in the accompanying website.

Box 41.12 provides the SPSS command sequence to the Kruskal-Wallis statistic.

For more than two *related* samples (e.g. the same group voting for three or more items, or the same grouping voting at three points in time), the Friedman test is applied. For example, in Tables 41.26 to 41.28 there are three variables ('The course encouraged and stimulated your motivation and willingness to learn'; 'The course encouraged you to take responsibility for your own learning'; and 'The teaching and learning tasks and activities consolidate learning through application'), all of which are voted on by the same group. The frequencies are given. Is there a statistically significant difference between the groups in their voting?

The Friedman test reports the mean rank and then the significance level; in the examples here the SPSS output has been reproduced in Tables 41.29 and 41.30.

One can see in Table 41.30 that, with a significance level of 0.838 (greater than 0.05), the voting by the same group on the three variables is not statistically significantly different, i.e. the null hypothesis is supported. The reporting of the results of the Friedman test can follow that of the Kruskal-Wallis test.

When the Friedman statistic was calculated to determine whether there was any statistically significant difference in the voting of the group on the three variables 'The course encouraged and stimulated your motivation and willingness to learn', 'The course encouraged you to take responsibility for your own learning', and 'The teaching and learning tasks and activities consolidate learning through application', no statistically significant difference was found between the group on the three variables in question ($\chi^2 = 0.353$, $\rho = 0.838$).

Box 41.13 provides the SPSS command sequence for the Friedman test.

For both the Kruskal-Wallis and the Friedman tests, as with the Mann-Whitney and Wilcoxon tests, *not* finding a statistically significant difference between groups can be just as important as finding a statistically significant difference between them, as the former suggests that nominal characteristics of the sample make no statistically significant difference to the voting, i.e. the voting is consistent, regardless of particular features of the sample. Similarly, as with the Mann-Whitney and the Wilcoxon tests, the Kruskal-Wallis and Friedman tests yield only statistical significance and not effect size.

BOX 41.12 SPSS COMMAND SEQUENCE FOR THE KRUSKAL-WALLIS STATISTIC

The SPSS command sequence for the Kruskal-Wallis statistic is: 'Analyze' → 'Nonparametric statistics' → 'Legacy Dialogs' → 'K Independent Samples' → Send the dependent variable to box 'Test variable list' and the independent variable to the 'Grouping variable' box → Click 'Define range' (which is activated when the 'Grouping variable' box contains the independent variable) and then type the number that you assigned to the first and last groups in the range in the SPSS file (e.g. School A: '1', School B: '2', School C: '3'). Ensure that the 'Kruskal-Wallis' box has been checked → Click 'OK'.

TABLE 41.26 FREQUENCIES FOR VARIABLE ONE IN THE FRIEDMAN TEST (SPSS OUTPUT)

the course encouraged and stimulated your motivation and willingness to learn

	Valid						Total
	not at all	very little	a little	quite a lot	a very great deal	Total	
Frequency	1	13	64	79	32	189	191
Valid Percent	.5	6.9	33.9	41.8	16.9	100.0	

TABLE 41.27 FREQUENCIES FOR VARIABLE TWO IN THE FRIEDMAN TEST (SPSS OUTPUT)

the course encouraged you to take responsibility for your own learning

	Valid						Total
	not at all	very little	a little	quite a lot	a very great deal	Total	
Frequency	1	9	64	85	30	189	191
Valid Percent	.5	4.8	33.9	45.0	15.9	100.0	

TABLE 41.28 FREQUENCIES FOR VARIABLE THREE IN THE FRIEDMAN TEST (SPSS OUTPUT)

the teaching and learning tasks and activities consolidate learning through application

	Valid				
	very little	a little	quite a lot	a very great deal	Total
Frequency	6	71	92	22	191
Valid Percent	3.1	37.2	48.2	11.5	100.0

TABLE 41.29 RANKINGS FOR THE FRIEDMAN TEST (SPSS OUTPUT)

Ranks

	Mean Rank
the course encouraged and stimulated your motivation and willingness to learn	1.98
the course encouraged you to take responsibility for your own learning	2.03
the teaching and learning tasks and activities consolidate learning through application	1.99

BOX 41.13 SPSS COMMAND SEQUENCE FOR THE FRIEDMAN TEST

The SPSS command sequence for the Friedman test is: 'Analyze' → 'Nonparametric statistics' → 'Legacy Dialogs' → 'K Related Samples' → Send to the 'Test variables' box the variables in which you are interested list' → Ensure that the 'Friedman' and the Kendall's W' boxes have been checked → Click 'OK'. The Kendall's W statistic yields a measure of effect size (see Chapter 39), though it is typically used to indicate the level of agreement ('concordance') between rankers rather than between variables (i.e. inter-rater reliability).

TABLE 41.30 SIGNIFICANCE LEVEL IN THE FRIEDMAN TEST (SPSS OUTPUT)

Test Statistics [a]

N	187
Chi-Square	.353
df	2
Asymp. Sig.	.838

a. Friedman Test

41.8 Conclusion

This chapter has introduced several inferential statistics and their related concepts:

- measures of difference for parametric data (t-test and ANOVA (one-way, two-way, Multiple Analysis of Variance) and post hoc tests of difference);
- measures of difference for non-parametric data (Mann-Whitney U test, Wilcoxon test, Kruskal-Wallis, Friedman);
- the chi-square test of independence and goodness of fit for univariate and bivariate categorical and ordinal variables, as a measure of difference between observed and expected values and as a test of association/difference;
- degrees of freedom.

It has indicated that, even with interval and ratio data, if the 'safety checks' indicate that they are unsuitable for parametric statistics, then non-parametric statistics should be used. The chapter has also included SPSS command sequences to run these statistics. The next chapter introduces more inferential statistics: regression analysis and standardization.

Companion Website

The companion website to the book provides additional material, data files and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Inferential statistics

CHAPTER 42

Regression analysis and standardization

This chapter introduces regression and multiple regression, and, arising from these, the need for standardized scores and how they can be calculated. The chapter proceeds thus:

- regression analysis (prediction tests for parametric data)
- simple linear regression (predicting the value of one variable from the known value of another variable)
- multiple regression (calculating the different weightings of independent variables on a dependent variable)
- standardized scores (used in calculating regressions and comparing sets of data with different means and standard deviations)

These statistics are powerful tools for analysing numerical data. We give several worked examples for clarification, and take the novice reader by the hand through these.

42.1 Regression analysis

Regression analysis enables the researcher to predict 'the specific value of one variable when we know or assume values of the other variable(s)' (Cohen and Holliday, 1996, p. 88). It is a way of modelling the relationship between variables. We concern ourselves here with simple linear regression and simple multiple regression, though we also reference stepwise multiple regression and logistic regression.

In using regression techniques, one has to be faithful to the assumptions underpinning them. Pallant (2016, chapters 13 and 14) and Tabachnick and Fidell (2013) set these out as follows, and they are 'safety checks' to ensure that the data are suitable for this set of statistical procedures.

Safety checks

For regression to be used safely there are several requirements:

- sample size: the larger, the better. Pallant (2016) suggests that fifteen cases for each independent variable are required, and that a formula can be applied to determine the minimum sample size required thus: sample size $\geq 50 + (8 \times$ number of independent variables), i.e. for ten independent variables one would require a minimum sample size of 130, i.e. $50 + (8 \times 10)$. Tabachnick and Fidell (2013) also suggest samples of $>50 + 8$ times the number of independent variables), and for stepwise regression (discussed below) there should be a minimum of forty cases for each independent variable;
- avoidance of multicollinearity, i.e. avoiding strong correlation ($r = 0.9$ or higher) between independent variables so that no independent variable is a perfect linear combination of another (avoid perfect 'multicollinearity');
- avoidance of singularity (where one variable is a combination of independent variables);
- avoidance of outliers (remove outliers);
- the measurements are from a random sample (or at least a probability-based one);
- all variables are real numbers (ratio data) (or at least the dependent variable must be);
- all variables are measured without error;
- there is an approximate linear (straight line) relationship between the dependent variable and the independent variable(s) (both individually and grouped);
- normal distribution of the variables;
- the residuals (explained below) for the dependent variable are approximately normally distributed and each value of the independent variables has equal and constant variance;
- homoscedasticity (the variance of the residuals for the dependent variable is the same); each residual is consistent across the range of values for all other variables;
- the residuals are not strongly correlated with the independent variables;
- for any two cases the correlation between the residuals should be zero (each case is independent of the others).
- interaction effects of independent variables are measured.

These safety checks should be applied prudently, as perfection is impossible. However, some of these are essential: random sampling; large sample; no multicollinearity or singularity; assumption of straight line linearity (rather than a curvilinear relationship); removal of outliers; ratio data; normal distributions of the residuals about the predicted dependent variable scores; homoscedasticity. SPSS easily runs tests for these, and we address this below.

42.2 Simple linear regression

In simple linear regression, the model includes one explanatory variable (the independent variable) and one explained variable (the dependent variable). For example, we may wish to see the effect of hours of study on levels of achievement in an examination, to see how much improvement can be predicted to be made to an examination mark from a given number of hours of study. 'Hours of study' is the independent variable and 'level of achievement' is the dependent variable. Conventionally, as in the example shown in Figure 42.1, one places the independent variable in the vertical axis and the dependent variable in the horizontal axis. In this example we have taken fifty cases of hours of study and student performance, and have constructed a scatterplot to show the distributions (SPSS performs this function at the click of two or three keys). We have also constructed a line of best fit (SPSS does this easily) to indicate the relationship between the two variables. The line of best fit is the closest straight line that can be constructed to take account of variance in the scores, and strives to have the same number of cases above it and below it and to make each point as close to the line as possible; for example, one can see that some scores are very close to the line and others are some distance away.

One can observe that the greater the number of hours spent in studying, generally the greater the level of achievement. This is akin to correlation. The line of

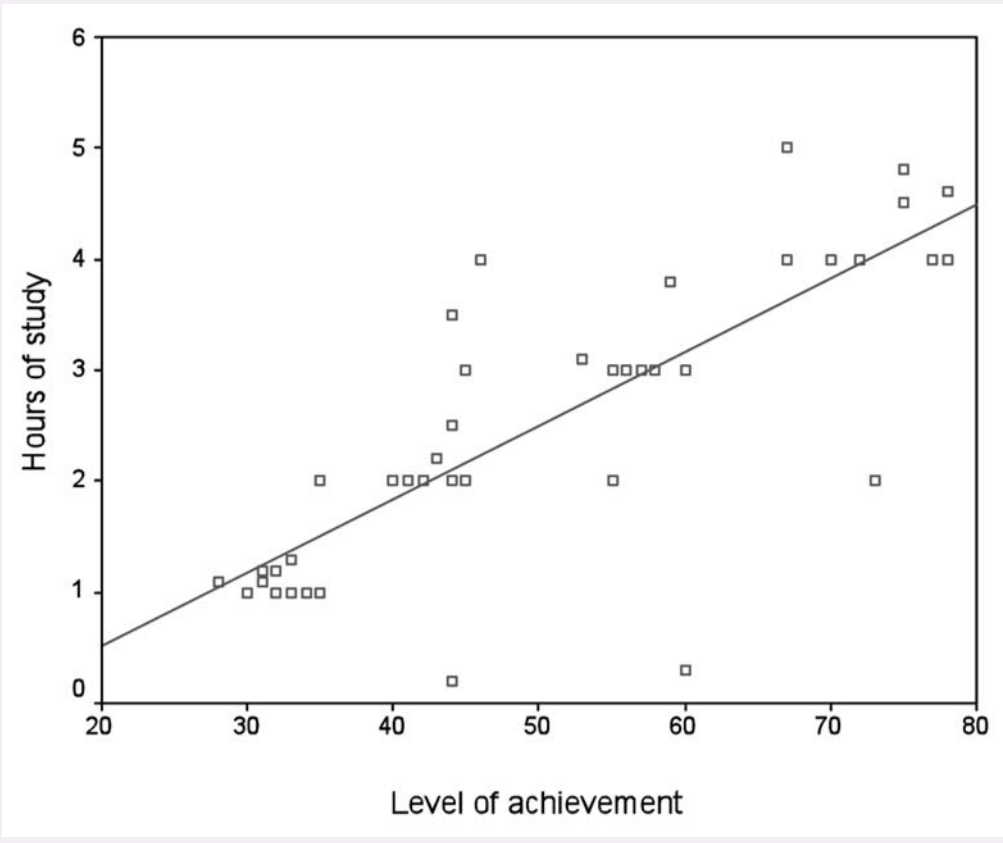


FIGURE 42.1 A scatterplot with the regression line (SPSS output)

best fit indicates not only that there is a positive relationship but that the slope of the line is quite steep. However, where regression departs from correlation is that regression provides an exact prediction of the value – the amount – of one variable when one knows the value of the other. One can read off the level of achievement, for example, if one were to study for two hours (43 marks out of 80) or for four hours (72 marks out of 80), of course, taking no account of variance. To help here, scatterplots (e.g. in SPSS) can insert grid lines.

It is dangerous to predict *outside* the limits of the line; simple regression is only used to calculate values *within* the limits of the actual line, and not beyond it. One can observe also that, though it is possible to construct a straight line of best fit (SPSS does this automatically), some of the data points lie close to the line and some lie a long way from the line; the distance of the data points from the line is termed the residuals, and this would have to be commented on in any analysis. A residual is the difference between the predicted and the actual score on the dependent variable (the distance from an actual score to the line of best fit). Ideally the residuals should be small, i.e. all the data points (values) on the graph should be close to the line of best fit (homoscedasticity), with few, if any, large exceptions (outliers or exceptional cases).

Where the line of best fit strikes the vertical axis is named the *intercept*. We return to this later, but at this stage we note that the line does not go through the origin (the 'zero') but starts a little way up the vertical line. In fact this is all calculated automatically by SPSS.

Let us look at a typical SPSS output shown in Table 42.1. This table provides the R square. The R square tells us how much variance in the dependent variable is explained by the independent variable in the calculation. First it gives us an R square value of 0.632, which indicates that 63.2 per cent of the variance is accounted for in the model, which is high. The 'Adjusted R square' is more accurate, and we advocate its use, as it automatically takes account of the number of independent variables. The Adjusted R square is usually smaller than the unadjusted R square, as it also takes account of the fact that one is looking at a sample rather than the whole population. Here the Adjusted R square is 0.625, and this shows that in the regression model that we have constructed, the independent variable accounts for 62.5 per cent of the variance in the dependent variable, which is high, i.e. our regression model is robust. Muijs (2004, p. 165) suggests that, for a goodness of fit from an Adjusted R square:

<0.1: poor fit
0.11–0.3: modest fit
0.31–0.5: moderate fit
>0.5: strong fit

SPSS then calculates the Analysis of Variance (ANOVA) (see Table 42.2). At this stage we will not go into all of the calculations here (typically SPSS prints out far more than researchers may need; for a discussion of df (degrees of freedom) we refer readers to Chapter 41). We go to the final column here, marked 'Sig.'; this is the significance level, and, because the significance is 0.000, we have a statistically significant relationship (stronger than 0.001) between the independent variable (hours of study) and the dependent variable (level of achievement), i.e. the relationship is not simply by chance.

All of this tells us that it is useful to proceed with the analysis. SPSS then gives us a table of coefficients, both unstandardized and standardized (Table 42.3). Here we advise researchers to opt for the standardized coefficients, the Beta weightings, as this gives greater precision, comparability and accuracy. The Beta weight (β) is the amount of standard deviation unit of change in the dependent variable for each standard deviation unit of change in the independent variable. In Table 42.3, the Beta weighting is 0.795; this tell us that, for every standard deviation unit change in the independent

TABLE 42.1 A SUMMARY OF THE R, R SQUARE AND ADJUSTED R SQUARE IN REGRESSION ANALYSIS (SPSS OUTPUT)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.795[a]	.632	.625	9.200

a. Predictors: (Constant), Hours of study

TABLE 42.2 SIGNIFICANCE LEVEL IN REGRESSION ANALYSIS (SPSS OUTPUT)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6988.208	1	6988.208	82.573	.000[a]
	Residual	4062.292	48	84.631		
	Total	11050.500	49			

a. Predictors: (Constant), Hours of study

b. Dependent Variable: Level of achievement

variable (hours of study), the dependent variable (level of achievement) will rise by 0.795 (79.5 per cent) of one standard deviation unit, i.e. in layperson's terms, for every one unit rise in the independent variable there is just over three-quarters of a unit rise in the dependent variable. This also explains why the slope of the line of best fit is steep but not quite 45 degrees – each unit of one is worth only 79.5 per cent of a unit of the other (Table 42.3).

Table 42.3 also indicates that the results are highly statistically significant (the 'Sig.' column (0.000) reports a significance level stronger than 0.001). Table 42.3 also includes a 'constant'; this is an indication of the vertical scale point where the line of best fit strikes the vertical axis, the intercept; the constant is sometimes taken out of subsequent analyses.

In reporting the example of regression, one could use a form of words thus:

> A scattergraph of the regression of hours of study on levels of achievement indicates a linear positive relationship between the two variables, with an Adjusted R square of 0.625. A standardized beta coefficient (β) of 0.795 is found for the variable 'hours of study', which is statistically significant ($\rho < 0.001$).

In simple regression the Beta (β) is the measure of effect size, as it is a correlation coefficient. The three main pieces of information to look for in a simple regression are (i) the Adjusted R square; (ii) the ANOVA significance level; and (iii) the Beta (β) value.

Box 42.1 presents the SPSS command sequence for simple regression.

42.3 Multiple regression

In linear regression we are able to calculate the effect of one independent variable on one dependent variable. However, it is often useful to be able to calculate the effects of two or more independent variables on a dependent variable. Multiple regression enables researchers to predict and weight the relationship between two or more *explanatory* – independent – variables and an *explained* – dependent – variable. We know from the previous example that the Beta (β) weighting gives an indication of how many standard deviation units will be changed in the dependent variable for each standard deviation unit of change in each of the independent variables. The Beta, as before, is the measure of effect size here.

TABLE 42.3 THE BETA COEFFICIENT IN A REGRESSION ANALYSIS (SPSS OUTPUT)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	26.322	2.982		8.828	.000
	Hours of study	9.567	1.053	.795	9.087	.000

a. Dependent Variable: Level of achievement

BOX 42.1 SPSS COMMAND SEQUENCE FOR SIMPLE REGRESSION

The SPSS commands for simple regression are in two stages. The first stage is to create a scatterplot, using these commands: 'Graphs' → 'Legacy Dialogs' → 'Scatter/Dot' → 'Simple Scatter' → 'Define' → Send over the dependent variable to the 'Y Axis' box and the independent variable to the 'X Axis' box → Click 'OK'. Edit the content by placing the cursor inside the scattergraph and right-clicking to create the SPSS command: 'Edit Content in separate window' → Click 'Add Fit line at Total' → Close the 'Edit' window. This creates the scattergraph. Then run the regression: 'Analyze' → 'Regression' → 'Linear' → Send the dependent variable to the 'Dependent' box and the independent variable to the 'Independent' box → Click 'Statistics' → Check the boxes 'Confidence Intervals', 'Select Descriptives', 'Part and Partial Correlations' and deselect 'Model Fit' → Click 'Continue' → Click 'OK'.

Let us take a worked example. An examination mark may be the outcome of study time and intelligence (Figure 42.2), and the formula here is:

Examination mark = β hours of study + β intelligence

Let us say that the β for hours of study is calculated by SPSS to be 0.65, and the β for intelligence is calculated to be 0.30. These are the relative weightings of the two independent variables. We wish to see how many marks in the examination a student will obtain who has an intelligence score of 110 and who studies for thirty hours per week. The formula becomes:

Examination mark
$= (0.65 \times 30) + (0.30 \times 110) = 19.5 + 33 = 52.5$

If the same student studies for forty hours then the examination mark could be predicted to be:

Examination mark
$= (0.65 \times 40) + (0.30 \times 110) = 26 + 33 = 59$

This enables the researcher to see the exact predicted effects of a particular independent variable on a dependent variable, when other independent variables are also present. In SPSS the constant is also calculated and this can be included in the analysis, to give the following, for example:

Examination mark = β hours of study + β intelligence + constant

Let us give an example with SPSS with more than two independent variables. Imagine that we wish to see how much improvement will be made to an examination mark from a given number of hours of study together with measured intelligence (e.g. IQ) and level of interest in the subject studied. We know from the previous example that the Beta weighting (β) gives us an indication of how many standard deviation units will be changed in the dependent variable for each standard deviation unit of change in each of the independent variables. The equation is:

Level of achievement in the examination = β hours of study + β IQ + β level of interest in the subject + constant

The constant is calculated automatically by SPSS. Each of the three independent variables – hours of study, IQ and level of interest in the subject – has its own Beta (β) weighting in relation to the dependent variable: level of achievement.

If we calculate the multiple regression using SPSS we obtain several tables of results (using fictitious data on fifty students) which we address here.

First, for Table 42.4, the Adjusted R square is very high indeed (0.975), indicating that 97.5 per cent of the variance in the dependent variable is explained by the independent variables, which is extremely high. Table 42.5 indicates that the Analysis of Variance is highly statistically significant (0.000), indicating that the relationship between the independent and dependent variables is very strong, i.e. not by chance.

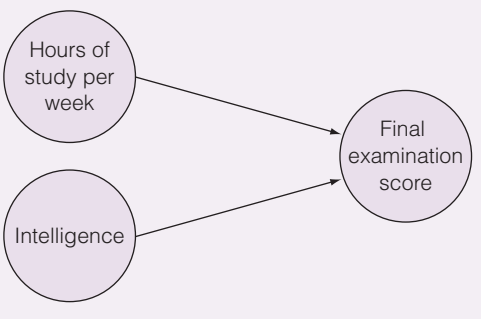


FIGURE 42.2 Multiple regression to determine relative weightings

TABLE 42.4 A SUMMARY OF THE R, R SQUARE AND ADJUSTED R SQUARE IN MULTIPLE REGRESSION ANALYSIS (SPSS OUTPUT)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.988[a]	.977	.975	2.032

a. Predictors: (Constant), Level of interest in the subject, Intelligence, Hours of study

TABLE 42.5 SIGNIFICANCE LEVEL IN MULTIPLE REGRESSION ANALYSIS (SPSS OUTPUT)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7969.607	3	2656.536	643.116	.000[a]
	Residual	190.013	46	4.131		
	Total	8159.620	49			

a. Predictors: (Constant), Level of interest in the subject, Intelligence, Hours of study

b. Dependent Variable: Level of achievement

The Beta (β) weighting of the three independent variables is given in the 'Standardized Coefficients' column of Table 42.6. The constant is given as 1.996.

It is important to note here that the Beta weightings for the three independent variables are calculated *relative to* each other rather than independent of each other. Hence we can say that, relative to each other:

- the independent variable 'hours of study' has the strongest positive effect on ($\beta = 0.920$) on the level of achievement, and this is statistically significant (the column 'Sig.' indicates that the level of significance, at 0.000, is stronger than 0.001);
- the independent variable 'intelligence' has a negative effect on the level of achievement ($\beta = -0.062$) but this is not statistically significant (at 0.644, $\rho > 0.05$);

TABLE 42.6 THE BETA COEFFICIENTS IN A MULTIPLE REGRESSION ANALYSIS (SPSS OUTPUT)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	21.304	10.675		1.996	.052
	Hours of study	9.637	1.863	.920	5.173	.000
	Intelligence	-6.20E-02	.133	-.062	-.466	.644
	Level of interest in the subject	.116	.135	.131	.858	.395

a. Dependent Variable: Level of achievement

- the independent variable 'level of interest in the subject' has a positive effect on the level of achievement $\beta = 0.131$), but this is not statistically significant (at 0.395, $\rho > 0.05$);
- the only independent variable that has a statistically significant effect (i.e. not by chance) on the level of achievement is 'hours of study'.

So, for example, with this knowledge, if we knew the hours of study, the IQ and the level of measured interest of a student, we could predict his or her expected level of achievement in the examination.

Box 42.2 provides the command sequence for multiple regression in SPSS.

Running the safety checks in multiple regression in SPSS

Safety checks here include checks for:

- sample size, random sampling and parametric data (can be checked before deciding whether to embark on multiple regression);
- collinearity;
- normality, linearity and homoscedasticity;
- outliers and distributions;
- residual scatterplot analysis.

We address below those areas which can only be conducted once the initial multiple analysis has been run (the last four bullet points above).

In the SPSS output that results from the command sequence, check for collinearity: (a) look at the table called 'Correlations', where the correlation coefficients between each independent variable and the dependent variable should be between 0.3 and 0.7; (b) look at the table called 'Coefficients' where the figures in the column labelled 'Tolerance' should be higher than 0.10 and the figures in the column labelled the 'Variance Inflation Factor' ('VIF') should be lower than 10. An example of this is given in Table 42.7, where the dependent variable 'Mathematics test score' is correlated with two independent variables 'How hard do you work for mathematics' and 'How many hours a week do you spend on your mathematics homework'. Here the Tolerance is 0.991 and the VIF is 1.009, both of which are 'safe'.

In the SPSS output, check for normality, linearity and homoscedasticity: (a) the points in the 'normal probability plot' (automatically produced by SPSS) should lie in a reasonably straight line running from the bottom left to the top right (linearity) and they should be close to the line of best fit (normality) and evenly distributed above and below the line (homoscedasticity) (Figure 42.3 gives an example of this).

In the output from SPSS, look at the scatterplot (automatically produced by SPSS). An example of this is given in Figure 42.4. Here the residuals are approximately rectangularly distributed with the centre of the box being in a straight line with the horizontal axis centre-point and the vertical axis centre point both going through zero.

In the SPSS output, check for outliers. Go to the table 'Residuals Statistics' (see Table 42.8 for an example of this). Go to the Mahalanobis Distance ('Mahal. Distance') (to identify outliers) to see if the figure is lower than the critical value; a table of critical values of chi-square is necessary here, and is available in most statistics textbooks or online. In the SPSS data file a new variable has been created in the data file (MAH_1). Go to the table of critical values of chi-square (provided in soft copy on the companion website), and use $\rho = 0.001$ as a significance level. In this instance there are two independent variables, for $\rho = 0.001$ the value is 13.82. But the 'Mahal. Distance' maximum here is 15.639, i.e. higher than the chi-square critical value; this is a problem. To solve this problem, go back to the Data Editor in SPSS, then Sort Cases, then sort by the new variable at the bottom of the data file (Mahalanobis Distance, MAH-1) in 'Descending' order. In the Data View window, the case with the

BOX 42.2 SPSS COMMAND SEQUENCE FOR MULTIPLE REGRESSION

To run multiple regression in SPSS, the command sequence is: 'Analyze' → 'Regression' → 'Linear'. Send over dependent variable to the 'Dependent' box. Send over independent variables to the 'Independent' box → Click 'Statistics'. Tick the boxes 'Estimates', 'Confidence Intervals', 'Model fit', 'Descriptives', 'Part and partial correlations', 'Collinearity diagnostics', 'Casewise diagnostics' and 'Outliers outside 3 standard deviations' → Click 'Continue' → Click 'Options' → Click 'Exclude cases pairwise' → Click 'Continue' → Click 'Plots'. Send over *ZRESID to the 'Y' box. Send over *ZPRED to the 'X' box → Click 'Normal probability plots' → Click 'Continue' → Click 'Save' → Click the 'Mahalanobis' box and the 'Cook's' box → Click 'Continue' → Click 'OK'. For further discussion of the SPSS commands and analysis of output we refer the reader to Pallant (2016, chapter 13) and Tabachnick and Fidell (2013, chapter 5).

TABLE 42.7 COEFFICIENTS TABLE FOR EXAMINING COLLINEARITY THROUGH TOLERANCE AND THE VARIANCE INFLATION FACTOR (VIF) (SPSS OUTPUT)

Coefficients[a]

Model		Coefficients[a]											
		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	23.517	3.941		5.968	0.000	15.775	31.260					
	How hard do you work for mathematics	4.777	0.283	0.576	16.867	0.000	4.221	5.334	0.601	0.603	0.573	0.991	1.009
	How many hours a week do you spend on your mathematics homework	2.811	0.375	0.256	7.504	0.000	2.075	3.547	0.312	0.319	0.255	0.991	1.009

Note
a Dependent Variable: Mathematics test score

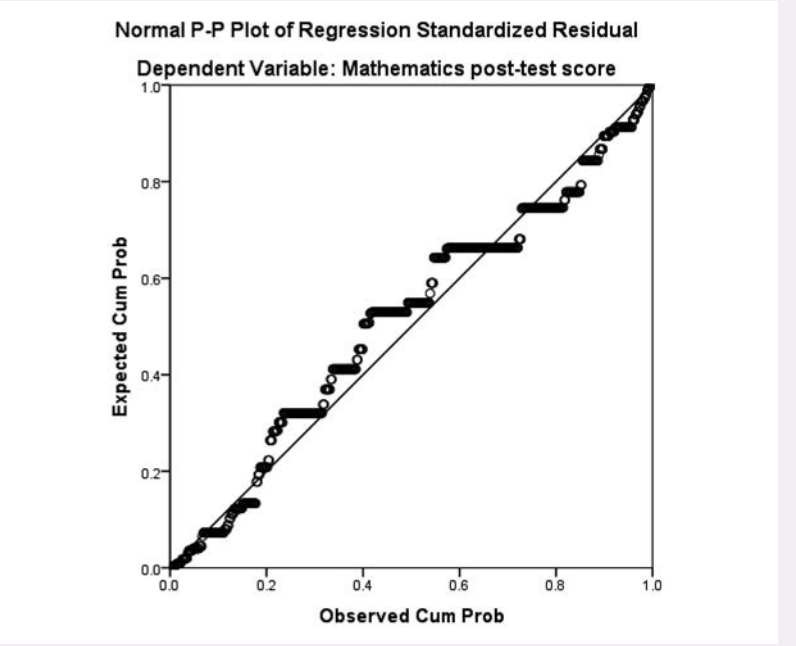


FIGURE 42.3 Normal probability plot for testing normality, linearity and homoscedasticity (SPSS output)

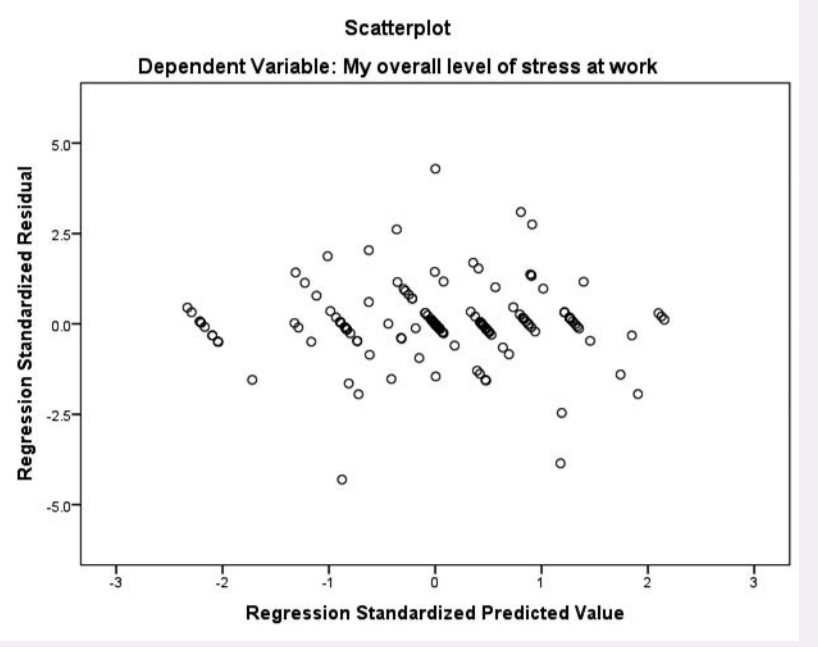


FIGURE 42.4 Scatterplot to check the distributions of the data (SPSS output, with horizontal and vertical lines added)

TABLE 42.8 CHECKING FOR OUTLIERS (SPSS OUTPUT)

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	65.11	99.40	90.68	6.973	500
Std. Predicted Value	–3.667	1.250	0.000	1.000	500
Standard Error of Predicted Value	0.382	1.481	00-.592	0.212	500
Adjusted Predicted Value	64.29	99.53	90.68	6.980	500
Residual	–31.810	25.333	0.000	8.096	500
Std. Residual	–3.921	3.123	0.000	0.998	500
Stud. Residual	–3.926	3.143	0.000	1.002	500
Deleted Residual	–31.881	25.706	0.005	8.155	500
Stud. Deleted Residual	–3.984	3.172	0.000	1.004	500
Mahal. Distance	0.111	15.639	1.996	2.569	500
Cook's Distance	0.000	0.107	0.002	0.008	500
Centered Leverage Value	0.000	0.031	0.004	0.005	500

Note
a Dependent Variable: Mathematics test score.

largest Mahalanobis Distance value will be at the top of the data file. If there are only a few then this is not a problem, but if there are many, or if the Mahalanobis Distance greatly exceeds the critical values, then consider removing those cases (there are only six cases here, so it is probably safe).

In the SPSS output, check for distributions and outliers in respect of the standardized residuals values: Table 42.9 provides the 'Casewise Diagnostics', i.e. the cases which are outside the normal range. Here cases number 9, 15, 22, 28, 133 and 350 in the SPSS file are outside the range and should be considered for removal. You can check to see if these outliers are exerting an undue influence on the results by going to the 'Cook's Distance' (Table 42.8); here the Cook's distance is 0.107, i.e. it is 'safe', as it should not exceed 1. If there is a problem then, as with the Mahalanobis Distance, go back to the SPSS data file and then sort cases in descending order, and then remove the 'offenders'.

In working with multiple regression then, using SPSS, there are six stages:

Stage 1: Conduct those 'safety checks' which can be conducted before the calculations proceed (e.g. random sampling, sample size).
Stage 2: Run the multiple regression.
Stage 3: Conduct the 'safety checks' once you have data (e.g. SPSS output): collinearity (correlation, Tolerance and VIF); normality; linearity; homoscedasticity; residuals scatterplot analysis (rectangular shape with centres of the rectangles on the

TABLE 42.9 CASEWISE DIAGNOSTICS (OUTLIER CASES) (SPSS OUTPUT)

Casewise diagnostics[a]

Case number	Std. Residual	Mathematics test score	Predicted value	Residual
9	–3.090	60	85.07	–25.066
15	3.019	100	75.51	24.489
22	3.068	90	65.11	24.888
28	3.123	100	74.67	25.333
133	–3.278	70	96.59	–26.588
350	–3.921	60	91.81	–31.810

Note
a Dependent Variable: Mathematics test score.

centre-points of the vertical and horizontal axes); outliers (remove them if necessary), with the Mahalanobis Distance and the Cook's Distance; standardized residuals values (Casewise diagnostics).

Stage 4: Note the Adjusted R square (to see the amount of explained variance that the independent variables have on the dependent variable). The R square is the multiple correlation coefficient squared. The Adjusted R square is the R square adjusted to take account of the sample size and the number of independent variables (it usually reduces the R square a little).

Stage 5: Check ANOVA and its significance level (to see if the model is statistically significant).

Stage 6: Note the Standardized Beta coefficients (β) and their statistical significance levels. The Standardized Beta Coefficient is the standardized regression coefficient. This tells you the amount of relative weight of each of the independent variables on the dependent variable so researchers can see which independent variable exerts more or less weight than the others. The standardized beta values indicate the number of standard deviations that scores in the dependent variable would change if there was one standard deviation unit change in the independent variable.

Multiple regression is useful in that it can take in a range of variables and enable researchers to calculate their relative weightings on a dependent variable. However, one has to be cautious: adding or removing variables affects their Beta coefficients. Morrison (2009, pp. 40–1) gives the example of Beta coefficients concerning the relative effects of independent variables on teacher stress (Table 42.10).

In Table 42.10 one can see the relative strengths (i.e. when one factor is considered in relation to the others included) of the possible causes of stress. It appears that 'teacher voice and support' exert the strongest influence on the outcome ('levels of stress') (beta of 0.323), followed by 'benefits and rewards' of teaching (beta of 0.205), then 'stress reproducing stress' (beta of 0.164) (i.e. the feeling of stress causes yet more stress), followed by 'burnout' (beta of 0.157), 'managing students' (beta of 0.116) and so on down the list. However, if we remove those variables connected with family ('family

TABLE 42.10 RELATIVE BETA WEIGHTINGS OF INDEPENDENT VARIABLES ON TEACHER STRESS (SPSS OUTPUT)

Beta Coefficients

	Standardized Coefficients	Significance level
	Beta	
Teacher voice and support	.323	.000
Workload	.080	.000
Benefits and rewards of teaching	.205	.000
Managing students	.116	.000
Challenge and debate	.087	.000
Family pressures	.076	.000
Considering leaving teaching	.067	.000
Emotions and coping	.044	.000
Burnout	.157	.000
Balancing work, family and cultural expectations	.100	.000
Local culture	.071	.000
Stress from family	.058	.000
Stress reproducing stress	.164	.000
Control and relationships	.092	.000

pressures', 'balancing work, family and cultural expectations' and 'stress from family') then the relative strengths of the remaining factors alter (see Table 42.11). In this revised situation (Table 42.11), the factor 'teacher voice and support' has slightly less weight, 'benefits and rewards of teaching' have added strength, and 'control and relationships' take on much greater strength.

On the other hand, if one adds in new independent variables ('principal behaviour' and 'clarity of jobs and goals') then the relative strengths of the variables alter again, as shown in Table 42.12. In this table the variable 'principal behaviour' greatly over-rides the other factors, and the order of the relative strengths of the other factors alters.

The point here is that the Beta coefficient (weightings) vary according to the independent variables included.

Further, variables may interact with each other and may be intercorrelated (the issue of multicollinearity), for example, Gorard (2001b, p. 172) suggests that poverty and ethnicity are likely to be correlated. If collinearity is discovered (e.g. if correlation coefficients between variables are higher than 0.80) then the researcher should consider removing one of the highly correlated variables, though caution has to be exercised here: the variable might be too important to remove, in which case Gorard advises researchers to create a single new variable that combines both previously intercorrelated variables.

SPSS automatically removes variables where there is strong covariance (collinearity). Muijs (2004) indicates that, in SPSS, one can find multicollinearity by looking at 'Collinearity Diagnostics' in the 'Statistics' command box of SPSS, and in the collinearity statistics, one should look at the 'Tolerance' column on the output. He indicates that values will vary from 0 to 1, and the higher the value the less is the collinearity, whereas a value close to 0 indicates that nearly all the variance in the variable is explained by the other variables in the model. For further discussion of collinearity, collinearity diagnostics and tolerance of collinearity, we refer the reader to Pallant (2016, chapter 13).

In reporting multiple regression, in addition to presenting tables (often of SPSS output), one can use a form of words thus, for example:

> Multiple regression was used, and the results include the adjusted R square (0.975), ANOVA ($\rho < 0.001$) and the standardized β coefficient of each component variable. Relative to each other, 'hours of study' exerted the greatest influence on level of achievement ($\beta = 0.920$, $\rho < 0.001$), 'level of interest' exerted a small and statistically insignificant influence on level of achievement ($\beta = 0.131$, $\rho = 0.395$), and 'intelligence' exerted a negative but statistically insignificant influence on level of achievement ($\beta = -0.062$, $\rho = 0.644$).

TABLE 42.11 ALTERED WEIGHTINGS IN BETA COEFFICIENTS (SPSS OUTPUT)

Beta Coefficients

	Standardized Coefficients	Significance level
	Beta	
Teacher voice and support	.316	.000
Workload	.096	.000
Benefits and rewards of teaching	.219	.000
Managing students	.114	.000
Challenge and debate	.099	.000
Considering leaving teaching	.102	.000
Emotions and coping	.091	.000
Burnout	.156	.000
Local culture	.131	.000
Stress reproducing stress	.162	.000
Control and relationships	.130	.000

TABLE 42.12 FURTHER ALTERED WEIGHTINGS IN BETA COEFFICIENTS (SPSS OUTPUT)

Beta Coefficients

	Standardized Coefficients	Significance level
	Beta	
Principal behaviour	.270	.000
Clarity of jobs and goals	.087	.000
Teacher voice and support	.154	.000
Workload	.109	.000
Benefits and rewards of teaching	.124	.000
Managing students	.102	.000
Challenge and debate	.095	.000
Family pressures	.071	.000
Considering leaving teaching	.081	.000
Emotions and coping	.084	.000
Burnout	.129	.000
Balancing work, family and cultural expectations	.098	.000
Local culture	.088	.000
Stress from family	.067	.000
Stress reproducing stress	.150	.000
Control and relationships	.086	.000

One variant of multiple regression is stepwise multiple regression. Here the computer enters variables one at a time, in a sequence, to see which adds to the explanatory power of a model, by looking at its impact on the R-squared – whether it increases the R-square value. This alternative way of entering variables and running the SPSS analysis in a 'stepwise' sequence is the same as above, except that in the 'Method' box the word 'Enter' should be replaced, in the drop-down box, with 'Stepwise'. In stepwise multiple regression the computer, not the researcher decides the order in which the independent variables are entered on the basis of significance testing and statistical computation (and this attracts criticism from some authors, e.g. Tabachnick and Fidell (2013) and Pallant (2016)). For working with stepwise regression we refer readers to these two sources.

Another type of multiple regression is logistic regression. It enables the researcher to work with categorical variables in a multiple regression where the dependent variable is a categorical variable with two or more values. Here the independent variables may be categorical, discrete or continuous. Logistic regression uses a Maximum Likelihood Estimation to produce a value between 0.0 and 1.0 which indicates the probability of the outcome. It does not require normality of distributions, but it is sensitive to outliers and collinearity.

Box 42.3 provides the command sequence for logistic regression in SPSS.

For more on logistic regression we refer the reader to Pallant (2016, chapter 14) and Tabachnick and Fidell (2013).

42.4 Standardized scores

Many forms of difference tests (see Chapter 41) and regression analysis with parametric data prefer to work with standardized scores, and we introduce these here. Imagine the following scenes:

1 Student (1) comes home from school and tells his parents that he scored a mark of 75 for a mathematics test; his parents berate him.

BOX 42.3 SPSS COMMAND SEQUENCE FOR LOGISTIC REGRESSION

To run logistic regression in SPSS, the command sequence is: 'Analyze' → 'Regression' → 'Binary Logistic' → Insert dependent variable in the 'Dependent' box → Insert independent variables into the 'Covariates' box → Click on 'Categorical' → Move your first categorical variable into the 'Categorical Covariates' box → Click the radio button 'First' → Click the 'Change' button → Repeat this for every categorical variable → Click 'Continue' to return to the first screen → Click 'Options' → Click the boxes 'Classification plots', 'Hosmer-Lemeshow goodness of fit', 'Casewise listing of residuals' and 'CI for Exp(B) → Click 'Continue' to return to the first screen → Click 'OK'.

2 Student (2) comes home from school and tells his parents that he scored a mark of 8 for a history test; his parents praise him.
3 Student (3) comes home from school and tells his parents that he scored a mark of 25 for an English test and a mark of 60 for a physics test; his parents praise him for both.
4 Student (4) comes home from school and tells his parents that he scored a mark of 80 for a geography test and a mark of 120 for a chemistry test; his parents berate him for both.

How can we explain these apparently discrepant behaviours? In the examples here we do not know the scales used, the range of scores, the means and the distributions around the means. For example, the student (1) scored 75 for his maths test and was berated because the mean score was 144 and the range was from 75 to 200, i.e. he scored very low on the test. On the other hand, student (2) scored 8 for his history test and was praised because that was the highest mark in the test, with an average mark of 4 out of a possible 10, and a range of 1 to 8. In the case of student (3) who was praised for scoring two very different marks (25 and 60), this was because the scales and range for the two tests varied, whereas student (4) who scored 80 for geography and 120 for chemistry was berated because both tests were marked out of 300 and the average mark for both was 220.

These examples show the need for researchers to compare like with like in using numerical data and scores. We need to know how to judge whether a mark is high or low and how to compare marks between one test and another. Therefore we need to know the *scale* of the marks, the *range* of the marks, the *mean* of the marks and the *distribution* of the marks either side of the mean. We need to know how to compare marks from a test which:

- uses one *scale* with marks from a test which uses another scale;
- has one *range* of marks with marks from a test that has another range of marks;
- has a *mean* which is different from the mean of another test;
- has a *distribution* around the mean which is different from the distribution of another test.

This is addressed by converting scores into standardized scores. Standardizing scores enables the researcher to judge whether a mark is high or low; it enables the researcher to compare marks between one test and another when two different tests have different scales, range, means and distributions around the mean. To standardize scores is to convert them into z-scores; z-scores have the same mean and standard deviation, even though the original sets of scores had different means and standard deviations, i.e. z-scores let researchers compare scores fairly. A z-score tells us how many standard deviations someone's scores lies above or below the mean. By standardizing different sets of scores (usually either a mean of zero and a standard deviation of one), this enables the researcher to compare like with like, to compare scores fairly.

To calculate the z-score we subtract the mean from the raw score and divide that answer by the standard deviation. The formula is thus:

$$z = \frac{\text{the actual score} - \text{the mean of the sample}}{\text{standard deviation of the sample}}$$

$$= \frac{x_i - \bar{x}}{s}$$

For example, if the raw (unadjusted) score is 15, the mean is 10 and the standard deviation is 4, then the standardized score is (15–10)/4=(5 ÷ 4)=1.25. Here the z-score tells us that the person's score is +1.25 standard deviations above the mean. However, we do not know whether this is a good score, a bad score, or, indeed, what it means. We need to see how this compares with other scores on the same distribution. Figure 42.5 plots the standardized scores on the normal curve of distribution, with the mean score of 0 (zero) and the standard deviation of 1, and marks the score of +1.25 on that diagram.

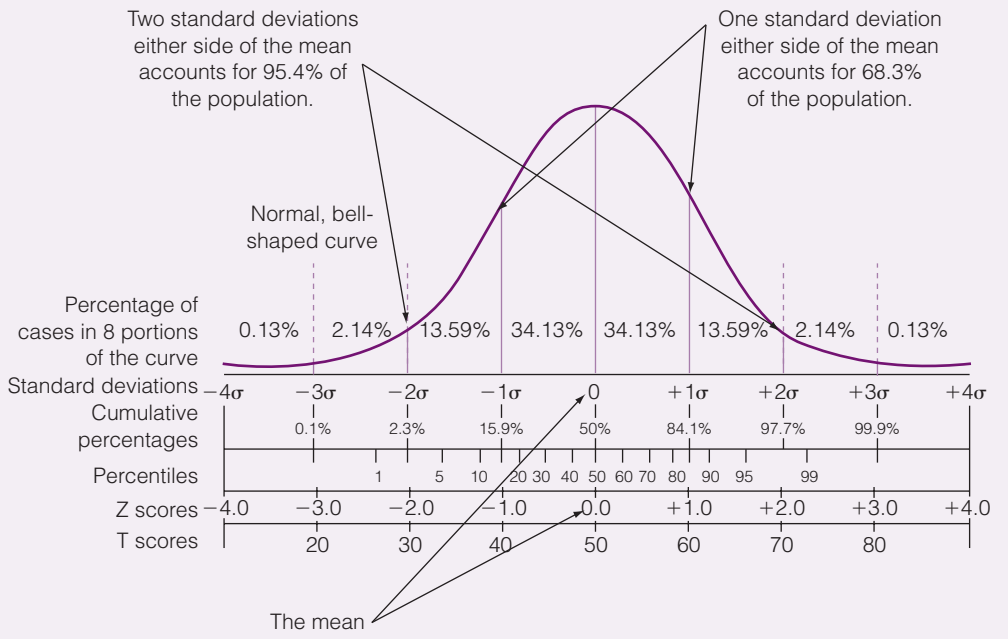


FIGURE 42.5 Standardizing scores

Looking at Figure 42.5, in our example, the person who scores +1.25 has scored very well. Had she scored 1 then she would have been better than 84.12 per cent of the population (34.13+34.13+13.59+2.14+0.13= 84.12): the percentage of people below her (see the lines marked 'Percentages of cases in 8 portions of the curve', 'Cumulative percentages' and 'Percentiles' in Figure 42.5). We know that she is higher than one standard deviation above the mean (she has scored 1.25, not 1), so we need to find where her score places her in terms of the rest of the population. For an exact indication of where she stands in relation to the rest of the population we can turn to statistical tables concerning 'areas under the normal curve' (on the Internet and in the appendices of most statistics books). Then we can simply read off the results (see Table 42.13 for an extract from such a table).

Referring to Table 42.13, she has a z-score of 1.25, so we go to the left-hand column, to the row marked '1.2'. Then we go to the column marked '0.05', as this gives us the second decimal place of the '1.25'. Then we see the value 0.3944 (emboldened and shaded), i.e. the person is 39.44 per cent above the mean of zero. We know from Figure 36.5 that 49.99 per cent of people are below zero (34.13+13.59+2.14+0.13=49.99, usually rounded to 50 per cent); now we add to that the 39.44 per cent above zero, giving a total of 89.44 per cent (from the rounded figure of 50 per cent). This tells us that, for the person with the z-score of 1.25, only 10.56 per cent (100 per cent minus 89.44 per cent) of the population is above her, so her score is very high.

Using the same table for another example, if a person receives a z-score of 1.56 then the table gives us a

TABLE 42.13 EXTRACT FROM AREA UNDER THE NORMAL CURVE OF DISTRIBUTION

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	**0.3944**	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441

reading of 0.4406, i.e. 44.06 per cent. We know from Figure 42.5 that 49.99 per cent of people are below zero (34.13+13.59+2.14+0.13=49.99, usually rounded to 50 per cent); now we add to that the 44.06 per cent above zero, giving a total of 94.06 per cent from the rounded figure of 50 per cent. This tells us that for the person with the z-score of 1.56 only 5.94 per cent (100 per cent minus 94.06 per cent) of the population is above that score, so the score is extremely high.

An online calculator of this is at:

- www.danielsoper.com/statcalc/calculator.aspx?id=2.

This calculator gives the cumulative area under the curve (a figure as a decimal fraction that is less than 1 (let us call it *X*). To find the area under the curve beyond that one point simply subtract this figure from 1 (the formula, then, is *1–X*) and, for a percentage, multiply it by 100. Another equally straightforward free online calculator of the area under the curve, and the position of a given z-score in that curve is given at:

- http://stattrek.com/online-calculator/normal.aspx.

Box 42.4 provides the SPSS command sequence for calculating standardized scores (z-scores).

Some people are uncomfortable with z-scores, as they do not like negative scores nor do they like an average being 0 (zero). To overcome this, z-scores can be converted to T-scores. To convert a z-score to a T-score, multiply the z-score by 10 and add 50 to the result. For example a z-score of 0.5, multiplied by 10 gives 5, and then, with 50 added, gives 55. The T-score is 55. Many IQ tests and standardized tests convert z-scores. For example, a common conversion in IQ tests is to multiply the z-score by 15 and add 100. So a z-score on an IQ test might be 0.5, multiplied by 15 gives 7.5, with 100 added gives 107.5, i.e. the IQ z-score converts to a T-score of 107.5.

Standardized scores are widely used in simple regression, as they enable researchers to compare different sets of scores on a fair basis.

42.5 Conclusion

This chapter has introduced several inferential statistics and their related concepts:

- simple regression and multiple regression (typical usage, stepwise regression and logistic regression);
- standardized scores and T-scores.

It has also included SPSS command sequences to run these statistics. Regression in all its forms is widely used in data analysis, and we commend it strongly to researchers. However, we also caution researchers to pay close attention to the several 'safety checks' before proceeding with regression analysis, together with confirmation that the assumptions underlying regression have been met. A widely used text on regression is Tabachnick and Fidell (2013).

BOX 42.4 SPSS COMMAND SEQUENCE FOR CALCULATING Z-SCORES

To calculate z-scores with SPSS, the command sequence is: 'Analyze' → 'Descriptive Statistics' → 'Descriptives' → 'Variables' → Click the box 'Save standardized values as variables' → Click 'OK'→ Two new variables will be created of the standardized scores.

Companion Website

The companion website to the book provides data files and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Factor analysis, cluster analysis and structural equation modelling

CHAPTER 43

This chapter addresses statistics and procedures for:

- conducting factor analysis
- what to look for in factor analysis output
- cluster analysis
- a note on structural equation modelling
- a note on multilevel modelling

Some of these items have significant coverage (e.g. factor analysis and cluster analysis), whilst others are more by way of introduction (e.g. structural equation modelling and multilevel modelling).

Techniques for grouping several research variables into factors are many and various: one of the best-known grouping techniques is factor analysis. Factor analysis is a widely used statistical technique in data analysis, and we introduce it in this chapter. We move to a note on cluster analysis as a way of organizing people/groups rather than variables, and then close with some introductory remarks on structural equation modelling and multilevel modelling.

43.1 Conducting factor analysis

Factor analysis is a method of grouping together variables which have something in common. It is a process which enables the researcher to take a set of variables and reduce them to a smaller number of underlying (latent) factors which account for as many variables as possible. It detects structures and commonalities in the relationships between variables. Thus it enables researchers to identify where different variables in fact are addressing the same underlying concept. For example, one variable could measure somebody's height in centimetres; another variable could measure the same person's height in inches; the underlying factor that unites both variables is height; it is a latent factor that is indicated by the two variables.

Factor analysis can take two main forms: *exploratory factor analysis* and *confirmatory factor analysis*. The former refers to the use of factor analysis (principal components analysis in particular) to explore previously unknown groupings of variables, to seek underlying patterns, clustering and groups. By contrast *confirmatory factor analysis* is more stringent, testing a found set of factors against a hypothesized model of groupings and relationships. Such a model derives from pre-established theory which informs the generation of the model, and the confirmatory factor analysis tests a theory of the latent processes and relationships. This section introduces a widely used kind of factor analysis: principal components analysis.

The analysis here uses SPSS output, as it is commonly used by educational researchers in undertaking principal components analysis.

As an example of factor analysis, one could have the following variables in a piece of educational research:

1. Student demotivation.
2. Poor student concentration.
3. Undue pressure on students.
4. Narrowing effect on curriculum.
5. Punishing the weaker students.
6. Overemphasis on memorization.
7. Testing only textbook knowledge.

These seven variables can be grouped together under the single overarching factor of 'negative effects of examinations'. Factor analysis, working through multiple correlations, is a method for grouping together several variables under one or more common factor(s).

To address factor analysis in more detail we provide a worked example. Consider the following variables concerning school effectiveness:

1. The clarity of the direction that is set by the school leadership.
2. The ability of the leader to motivate and inspire the educators.
3. The drive and confidence of the leader.
4. The consultation abilities/activities of the leader.
5. The example set by the leader.
6. The commitment of the leader to the school.
7. The versatility of the leader's styles.
8. The ability of the leader to communicate clear, individualized expectations.

9 The respect in which the leader is held by staff.
10 The staff's confidence in the Senior Management Team.
11 The effectiveness of the teamwork of the Senior Management Team.
12 The extent to which the vision for the school impacts on practice.
13 Educators given opportunities to take on leadership roles.
14 The creativity of the Senior Management Team.
15 Problem-posing, problem-identifying and problem-solving capacity of the Senior Management Team.
16 The use of data to inform planning and school development.
17 Valuing of professional development in the school.
18 Staff consulted about key decisions.
19 The encouragement and support for innovativeness and creativity.
20 Everybody is free to make suggestions to inform decision-making.
21 The school works in partnership with parents.
22 People take positive risks for the good of the school and its development.
23 Staff voluntarily taking on coordination roles.
24 Teamwork among school staff.

Here we have twenty-four different variables and the question which might concern researchers here is 'are there any underlying groups of factors' ('latent variables') that can embrace several of these variables, or of which the several variables are elements or indicators? Factor analysis indicates whether there are. In what follows we distinguish *factors* from *variables*; a *factor* is an underlying or latent feature in which groups of variables are included; a *variable* is one of the elements that can be a member of an underlying factor. In our example here we have twenty-four variables and, as we shall see, five factors.

Let us imagine that we have gathered data from 1,000 teachers in several different schools, and we wish to see how the twenty-four variables above can be grouped, based on the teachers' voting (using ratio data by awarding marks out of ten for each of the variables). (This follows the rule that there should be more subjects in the sample than there are variables.)

In what follows we set out a five-stage model in conducting factor analysis:

Stage 1: safety checks;
Stage 2: data processing and initial analysis;
Stage 3: constructing the factors from the variables;
Stage 4: naming the factors;
Stage 5: reporting the factor analysis.

This takes researchers from setting up the factor analysis to conducting the processing and analysis, to reporting the results. We also provide readers with the SPSS command sequence for Principal Components Analysis (PCA).

Factor analysis (sometimes termed 'principal axis factoring') is not the same as PCA, even though the terms are frequently used interchangeably. A major difference between the two is that in PCA all the variance in the data is analysed whereas in factor analysis only the shared variance is analysed, thereby excluding unique variance (Dancey and Reidy, 2011, p. 457). Further, PCA is used for reducing a large set of variables into factors, whereas factor analysis is used in causal modelling, particularly in confirmatory factor analysis where a hypothesized model or theory of relationships is tested (p. 457). In this chapter we focus on PCA.

Stage 1: safety checks

The first stage in factor analysis is to conduct 'safety checks' to see if the data are suitable for factor analysis, and to check whether the assumptions underpinning factor analysis have been met. There are several assumptions that factor analysis makes, and these must be addressed fairly in deciding whether, in fact, it is safe to proceed with factor analysis (cf. Tabachnick and Fidell, 2013). Factor analysis uses correlations, and several of the comments below concern correlations (see Chapter 40 for a discussion of correlations).

- *Sample size.* The suggested sample size varies in the literature, from a minimum of 30 to a minimum of 300. Tabachnick and Fidell (2013) suggest that a sample size of 50 is very poor, 100 is poor, 200 is fair, 300 is good, 500 is very good and 1,000 is excellent; they suggest that 300 should be regarded as a general minimum, and if the sample size is small then the factor loadings (discussed later) should be high. Some authors suggest a minimum of 10 cases per variable, whilst others suggest a minimum of between 150 and 200 cases in total, regardless of the number of variables. Bryman and Cramer (1990, p. 255) suggest no fewer than 100 subjects in the total sample.
- *Number of variables.* It is important to have neither too few nor too many variables: too few and the extraction of the factors may only extract one or two variables per factor, and this gives very little 'added value'. Too many and the number of factors extracted could be so many as to be unhelpful in identifying underlying latent factors.

- *Ratio of sample size to number of variables.* Different ratios are given in literature, from 5:1 to 30:1.
- *Interval and ratio data*. Ordinal data may also be used if this does not distort the underlying metric scaling.
- *Sampling adequacy.* The Kaiser-Mayer-Olkin (KMO) measure of sampling adequacy, which correlates pairs of variables and the magnitude of partial correlations among variables and which requires many pairs of variables to be statistically significantly correlated, should yield an overall measure of 0.6 or higher (maximum is 1). If the KMO index is high (1.0) then Principal Components Analysis can be conducted; if it is low (around 0.0) then PCA is not relevant.
- *Intercorrelations between variables.* Correlation coefficients (see Chapter 40) between variables should be no less than 0.3, as below this the data may not be suitable for finding latent, underlying factors, as the variables are not sufficiently closely related. Factor analysis with only low intercorrelations between variables (indicated in a correlation matrix, discussed below) is likely to generate as many factors as there are original variables, and this defeats the original purpose of factor analysis which is to reduce data and to generate clusters containing several variables each. Above 0.6 and there may be problems of multicollinearity (see Chapter 41), and this can be identified in a correlation matrix, which SPSS automatically calculates, and variables removed before conducting the factor analysis. The Tolerance and Variation Inflation Factor (VIF) (see Chapter 42) screen for collinearity, though typically the KMO statistic, which conducts a series of partial correlations, can also be used here, and SPSS can run this automatically; this enables the distinctiveness (unrelatedness) of each factor to be assured. The Bartlett test of sphericity, which investigates the correlations between variables, should show statistical significance ($\rho < 0.05$) (mainly used where the number of cases per variable is five or fewer).
- *Intercorrelations between factors.* The factors should not be highly correlated with each other (the principle of orthogonality).
- *Normal distributions.* Factor analysis assumes a normal distribution (measured by kurtosis and skewness). It is particularly important to screen the data for normality if the sample is small.
- *Linearity.* Principal Components Analysis (PCA) assumes linear (straight line) relationships between variables rather than, for example, curvilinearity (see Chapter 40). It is important to screen the data for linearity, particularly if the sample size is small.
- *Outliers.* The presence of outliers can distort calculations in factor analysis. Researchers can use the Mahalanobis Distance calculation (see Chapter 42) to identify cases which are outliers and then remove them before conducting the factor analysis.
- *Selection bias/proper specification.* Excluding relevant variables and including irrelevant variables can affect the nature of the factors extracted. Researchers must decide carefully which variables to include and exclude in conducting factor analysis.
- *Theoretical underpinning of factors.* Factor analysis assumes that there are underlying dimensions which are shared by clusters of variables in order to extract a factor. This refers to the need for there to be a strong theoretical underpinning of the factors. Factor analysis cannot create valid factors if none exist in the original data; even though factor analysis can create factors, if they have little or no theoretical substance then they are likely to be worthless. Factors and their labels must have face validity and strong theoretical grounding.

If the data are suitable for factor analysis then the researcher can proceed.

Stage 2: data processing and initial analysis

The analysis in the example here assumes that the data are suitable for factor analysis to proceed and the material discussed below is based on SPSS processing and output. At first SPSS produces a correlation matrix so that the researcher can check the intercorrelations mentioned above. Then it produces a table of extracted factors (Table 43.1).

Though Table 43.1 seems to contain a lot of complicated data, in fact most of this need not trouble us at all. SPSS has automatically found and reported five factors for us through correlational analysis, and it presents data on these five factors (the first five rows of the chart, marked 'Component'). Table 43.1 takes the twenty-four variables (listed in order on the left hand column (Component)) and then it provides three sets of readings: Eigenvalues, Extraction Sums of Squared Loadings, and Rotation Sums of Squared Loadings. Eigenvalues are measures of the variance between factors, and are the sum of the squared loadings for a factor, representing the amount of variance accounted for by that factor. We are only interested in those Eigenvalues that are greater than 1, since those that are smaller than 1 generally are not of interest to researchers as they account for less than the variation explained by a single variable. Indeed SPSS automatically filters out for us the Eigenvalues that are greater than 1, using

TABLE 43.1 INITIAL SPSS OUTPUT FOR PRINCIPAL COMPONENTS ANALYSIS (SPSS OUTPUT)

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	9.343	38.930	38.930	9.343	38.930	38.930	4.037	16.820	16.820
2	1.424	5.931	44.862	1.424	5.931	44.862	2.810	11.706	28.527
3	1.339	5.580	50.442	1.339	5.580	50.442	2.779	11.578	40.105
4	1.220	5.085	55.526	1.220	5.085	55.526	2.733	11.386	51.491
5	1.085	4.520	60.047	1.085	4.520	60.047	2.053	8.556	60.047
6	.918	3.825	63.872						
7	.826	3.443	67.315						
8	.723	3.013	70.329						
9	.685	2.855	73.184						
10	.658	2.743	75.927						
11	.623	2.596	78.523						
12	.562	2.342	80.864						
13	.532	2.216	83.080						
14	.512	2.132	85.213						
15	.493	2.055	87.268						
16	.466	1.942	89.210						
17	.437	1.822	91.032						
18	.396	1.650	92.682						
19	.376	1.566	94.247						
20	.364	1.517	95.764						
21	.307	1.280	97.044						
22	.271	1.129	98.174						
23	.232	.965	99.138						
24	.207	.862	100.000						

Extraction Method: Principal Component Analysis.

the Kaiser criterion (in SPSS this is termed the Kaiser Normalization).

A scree plot can also be used at this stage, to identify and comment on factors (this is available at a keystroke in SPSS). A scree plot shows each factor on a chart, in descending order of magnitude of amount of variance explained. For researchers, the scree plot becomes interesting where it flattens out (like the rubble that collects at the foot of a scree), as this indicates very clearly which factors account for a lot of the variance, and which account for little. In the scree plot here shown in Figure 43.1 one can see that the scree flattens out considerably after the first factor, and then it levels out a little for the next four factors, tailing downwards all the time. This suggests that the first factor is the significant factor in explaining the greatest amount of variance.

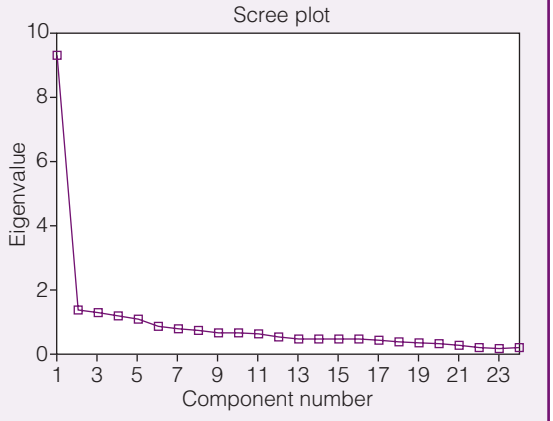


FIGURE 43.1 A scree plot (SPSS output)

In using the scree plot one can look for the 'bend in the elbow' of the data (after factor one), and then regard those factors above the bend in the elbow as being worthy of inclusion, and those below the bend in the elbow as being relatively unimportant. However, this is draconian, as it risks placing too much importance on those items above the bend in the elbow and too little importance on those below it. The scree plot adds little to the variance table presented above in Table 43.1, though it does enable one to see at a glance which are the significant and less important factors, or, indeed which factors to focus on (the ones before the scree levels off) and which to ignore.

Next we turn to the columns labelled 'Extraction Sums of Squared Loadings'. The Extraction Sums of Squared Loadings contain two important pieces of information. First, in the column marked '% of Variance' SPSS tells us how much variance is explained by each of the factors identified, in order from the greatest amount of variance to the least amount of variance. So, here the first factor accounts for 38.930% of the variance in the total scenario – a very large amount – whilst the second factor identified accounts for only 5.931 per cent of the total variance, a much lower amount of explanatory power. By showing us how much variance in the total picture is explained by each factor we can see which factors possess the most and least explanatory power – the power to explain the total scenario of twenty-four factors. Secondly, SPSS keeps a score of the cumulative amount of explanatory power of the five factors identified. In the column 'Cumulative' it tells us that in total 60.047 per cent of the total picture (of the twenty-four variables) is accounted for – explained – by the five factors identified. This is a moderate amount of explanatory power, and researchers would be happy with this.

The three columns under 'Extraction Sums of Squared Loadings' give us the initial, rather crude, unadjusted percentage of variance of the total picture explained by the five factors found. These are crude in the sense that the full potential of factor analysis has not been caught. What SPSS has done here is to plot the factors on a two-dimensional chart (which it does not present in the data output) to identify groupings of variables, the two dimensions being vertical and horizontal axes as in a conventional graph such as a scattergraph. On such a two-dimensional chart some of the factors and variables could be plotted quite close to each other, such that discrimination between the factors would not be very clear. However, if we were to plot the factors and variables on a three-dimensional chart that includes not only horizontal and vertical axes but also *depth* by *rotating* the plotted points through ninety degrees, then the effect of this would be to bring closer together those variables that are similar to each other and to separate them more fully – in distance – from those variables that have no similarity to them, i.e. to render each group of variables (factors) more homogeneous and to separate more clearly one group of variables (factor) from another group of variables (factor). The process of rotation keeps together those variables that are closely interrelated and keeps them apart from those variables that are not closely related. This is represented in Figure 43.2.

This distinguishes more clearly one factor from another than that undertaken in the Extraction Sums of Squared Loadings.

Rotation can be conducted in many ways (Tabachnick and Fidell, 2013), of which there are two main forms:

- Direct Oblimin: which is used if the researcher believes that there may be correlations between the *factors* (an oblique, correlated) rotation;
- Varimax rotation: which is used if the researcher believes that the *factors* may be uncorrelated (orthogonal).

Pallant (2016) argues for the importance of researchers giving strong consideration to the Direct Oblimin rotation. Even though it is more difficult to interpret, it is often actually more faithful to the correlated nature of the data and factors. The default setting in SPSS is the orthogonal, Varimax rotation, and this may misrepresent existing correlations between the factors, even though it is easier to analyse. Indeed Pallant (2016) suggests starting with Direct Oblimin rotation.

Rotation in the example here is undertaken by *Varimax rotation*. This maximizes the variance between factors and hence helps to distinguish them from each other. In SPSS the rotation is called *orthogonal* because

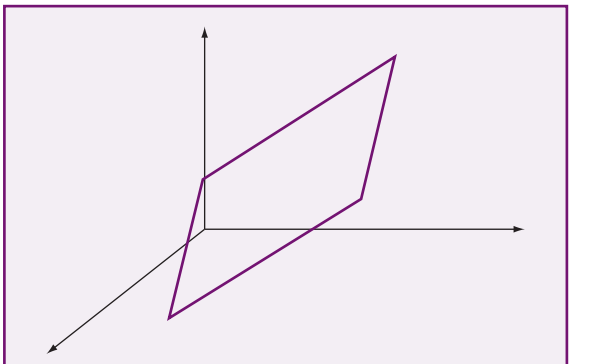


FIGURE 43.2 Three dimensional rotation

the factors are unrelated to, and independent of, each other.

In the column 'Rotation Sums of Squared Loadings' the fuller power of factor analysis is tapped, in that the rotation of the variables from a two-dimensional to a three-dimensional chart has been undertaken, thereby identifying more clearly the groupings of variables into factors, and separating each factor from the other much more clearly. We advise researchers to use the Rotation Sums of Squared Loadings rather than the Extraction Sums of Squared Loadings. With the Rotation Sums of Squared Loadings the percentage of variance explained by each factor is altered, even though the total cumulative percentage (60.047 per cent) remains the same. For example, the first factor in the rotated solution no longer accounts for 38.930 per cent as in the Extraction Sums of Squared Loadings, but only 16.820 per cent of the variance; factors 2, 3 and 4, which each only accounted for just over 5 per cent of the variance in the Extraction Sums of Squared Loadings now each account for over 11 per cent of the variance; and factor 5, which accounted for 4.520 per cent of the variance in the Extraction Sums of Squared Loadings now accounts for 8.556 per cent of the variance in the Rotation Sums of Squared Loadings.

By the end of this second stage we can see that:

1. factor analysis brings variables together into homogeneous and distinct groups, each of which is a factor and each of which has an Eigenvalue of greater than 1;
2. factor analysis in SPSS indicates the amount of variance in the total scenario explained by each individual factor and all the factors together (the cumulative percentage);
3. the Rotation Sums of Squared Loadings is preferable to the Extraction Sums of Squared Loadings.

We are ready to proceed to the third stage.

Stage 3: constructing the factors from the variables

Stage 3 consists of presenting a matrix of all of the relevant data and variables for the researcher to identify which variables belong to which factor (Table 43.2). SPSS presents what at first sight is a bewildering set of data, but the reader is advised to keep cool and to look at the data slowly, as, in fact, they are not complicated. SPSS often presents researchers with more data than they need, overwhelming the researcher with data. In fact the data in Table 43.2 are comparatively straightforward.

Across the top of the matrix in Table 43.2 we have a column for each of the five factors (1–5) that SPSS has found for us. The left-hand column prints the names of each of the twenty-four variables with which we are working. We can ignore those pieces of data which contain the letter 'E' (exponential), as these contain figures that are so small as to be able to be discarded. Look at the column labelled '1' (factor 1). Here we have a range of numbers, from 0.114 (for the variable 'Teamwork amongst school staff') to 0.758 (for the variable 'The drive and confidence of the leader'). The researcher now has to use her professional judgement to decide what the 'cut-off' points should be for inclusion in the factor. Not all twenty-four variables will appear in factor 1, only those with high values (factor loadings – the amount that each variable contributes to the factor in question). The decision on which variables to include in factor 1 is not a statistical matter but a matter of professional judgement, informed by theory and judgements about whether the variables 'hang together' – cluster – in a single factor. Factor analysis is an art as well as a science. The researcher has to find those variables with the highest values (factor loadings) and include those in the factor. The variables chosen should not only have high values but also have values that are conceptually and numerically close to each other (homogeneous) and which are some numerical distance away from the other variables. In the column labelled '1' we can see that there are seven such variables, and we set these out in the example below. Other variables from the list are some numerical distance away from the variables selected (see below) and also seem to be conceptually unrelated to the seven variables identified for inclusion in the factor. The variables selected are high, close to each other both numerically and conceptually, and distant from the other variables. The lowest of these seven values (factor loadings) is 0.513; hence the researcher would report that seven variables had been selected for inclusion in factor 1, and that the cut-off point was 0.51 (i.e. the lowest point, above which the variables have been selected). Having such a high cut-off point gives considerable power to the factor. Hence we have factor 1, which contains seven variables.

Let us look at a second example, that of factor 2 (the column labelled '2' in Table 43.2). Here we can identify four variables that have high values that are close to each other and yet are some numerical distance away from the other variables. These four variables constitute factor 2, with a reported cut-off point of 0.445. At first glance it may seem that 0.445 is low; however, recalling that the data in the example were derived from 1,000 teachers, 0.445 is still highly statistically significant, statistical significance being a combination of the coefficient *and* the sample size.

TABLE 43.2 THE ROTATED COMPONENTS MATRIX IN PRINCIPAL COMPONENTS ANALYSIS (SPSS OUTPUT)

Rotated Component Matrix[a]

	Component				
	1	2	3	4	5
The clarity of the direction that is set by the school leadership	.559	.133	7.552E-02	.248	.212
The ability of the leader to motivate and inspire the educators	.743	.142	.176	9.058E-02	.160
The drive and confidence of the leader	.758	2.151E-02	.122	2.796E-02	.222
The consultation abilities/activities of the leader	.548	.342	.208	.278	.160
The example set by the leader	.572	.239	.126	.319	.209
The commitment of the leader to the school	.513	.290	.252	.329	.137
The versatility of the leader's styles	.284	.332	.377	.285	5.668E-02
The ability of the leader to communicate clear, individualized expectations	.449	.246	.303	.351	.205
The respect in which the leader is held by staff	.184	7.988E-02	.154	.810	.240
The staff's confidence in the SMT	.180	.121	7.859E-02	.809	.279
The effectiveness of the teamwork of the SMT	.385	.445	.249	.443	8.104E-02
The extent to which the vision for the school impacts on practice	.413	.341	.305	.379	.113
Educators given opportunities to take on leadership roles	.247	.225	.494	.339	-2.66E-02
The creativity of the SMT	.212	7.188E-02	.822	-2.97E-03	.189
Problem-posing, problem-identifying and problem-solving capacity of SMT	.459	.351	.262	.361	-3.21E-02
The use of data to inform planning and school development	.690	.167	.188	5.158E-02	-3.79E-02
Valuing of professional development in the school	.187	.249	.551	.260	7.013E-02
Staff consulted about key decisions	.148	6.670E-02	.854	7.531E-02	.167
The encouragement and support for innovativeness and creativity	.143	5.187E-02	.189	.269	.661
Everybody is free to make suggestions to inform decision-making	.165	.150	.172	.264	.642
The school works in partnership with parents	.222	.804	8.173E-02	.143	.199
People take positive risks for the good of the school and its development	.206	.778	8.998E-02	.181	2.635E-02
Staff voluntarily taking on coordination roles	.195	.210	2.681E-02	3.660E-02	.779
Teamwork amongst school staff	.114	.642	.220	-3.41E-02	.277

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

We repeat this analysis for all five factors, deciding the cut-off point and which variables to include, looking for homogeneous high values and numerical and conceptual distance from other variables in the list.

Stage 4: naming the factors

By this time we have identified five factors. However, neither SPSS nor any other software package tells us what to name each factor. The researcher has to devise a name that describes the factor in question. This can be challenging, as it has to catch the issue that is addressed by all the variables that are included in the factor. We have done this for all five factors, and we report this below, with the factor loadings for each variable reported in brackets.

Factor 1: Leadership skills in school management
Cut-off point: 0.51
Variables included:

- The drive and confidence of the leader (factor loading 0.758).
- The ability of the leader to motivate and inspire the educators (factor loading 0.743).
- The use of data to inform planning and school development (factor loading 0.690).
- The example set by the leader (factor loading 0.572).
- The clarity of the direction set by the school leadership (factor loading 0.559).

- The consultation abilities/activities of the leader (factor loading 0.548).
- The commitment of the leader to the school (factor loading 0.513).

Factor 2: Parent and teacher partnerships in school development
Cut-off point: 0.44
Variables included:

- The school works in partnership with parents (factor loading 0.804);
- People take positive risks for the good of the school and its development (factor loading 0.778);
- Teamwork amongst school staff (factor loading 0.642);
- The effectiveness of the teamwork of the SMT (factor loading 0.445).

Factor 3: Promoting staff development by creativity and consultation
Cut-off point: 0.55
Variables included:

- Staff consulted about key decisions (factor loading 0.854).
- The creativity of the smt (senior management team) (factor loading 0.822).
- Valuing of professional development in the school (factor loading 0.551).

Factor 4: Respect for, and confidence in, the senior management
Cut-off point: 0.44
Variables included:

- The respect in which the leader is held by staff (factor loading 0.810).
- The staff's confidence in the SMT (factor loading 0.809).
- The effectiveness of the teamwork of the SMT (factor loading 0.443).

Factor 5: Encouraging staff development through participation in decision making
Cut-off point 0.64
Variables included:

- Staff voluntarily taking on coordination roles (factor loading 0.779).
- The encouragement and support for innovativeness and creativity (factor loading 0.661).
- Everybody is free to make suggestions to inform decision making (factor loading 0.642).

Each factor should usually contain a minimum of three variables, though this is a rule of thumb rather than a statistical necessity. Further, in the example here, though some of the variables included have considerably lower factor loadings than others in that factor (e.g. in factor 2 the variable 'the effectiveness of the teamwork of the SMT' (0.445)), nevertheless the conceptual similarity of this to the other variables in that factor, coupled with the fact that, with 1,000 teachers in the study, 0.445 is still highly statistically significant, combine to suggest that this still merits inclusion. As mentioned earlier, factor analysis is an art as well as a science.

If one wished to suggest a more stringent level of exactitude then a higher cut-off point could be taken. In the example above, factor 1 could have a cut-off point of 0.74, thereby including only two variables in the factor; factor 2 could have a cut-off point of 0.77, thereby including only two variables in the factor; factor 3 could have a cut-off point of 0.82, thereby including only two variables in the factor; factor 4 could have a cut-off point of 0.80, thereby including only two variables in the factor; and factor 5 could have a cut-off point of 0.77, thereby including only one variable in the factor. The decision on where to place the cut-off point is a matter of professional judgement when reviewing the data, but it makes little sense to have a factor containing only one or two variables.

Stage 5: reporting the factor analysis

In reporting factor analysis the following points should be included, for example:

- the kind of extraction method used (e.g. Principal Components Analysis) and why;
- the kind of rotation used (e.g. Varimax) and why;
- the use of Eigenvalues, KMO and Bartlett tests and what they show;
- the total amount of variance explained and whether this is high, medium or low;
- the amount of explained variance of each factor and whether it is high, medium or low;
- the cut-off points in the factor loadings for each variable and why;
- the titles given to each factor;
- which factors have the highest and lowest explained variance, and what this shows.

The reporting should also draw attention to specific points in relevant tables of data. It might also include indications of which variables were included in each factor and why, their factor loadings (though a table would normally include these), and which variables were excluded from

all/any of the factors and why. A short introductory commentary could be provided, for example:

> In order to obtain conceptually similar and significant clusters of issues of the variables, principal components analysis with Varimax rotation and Kaiser Normalization were conducted as the factors were deemed to be orthogonal. Eigenvalues equal to or greater than 1.00 were extracted and the Kaiser-Meyer-Olkin (0.845) and Bartlett tests of sphericity ($\rho = 0.000$) indicated that the data were suitable for factorization. From the twenty-four variables included, orthogonal rotation of the variables yielded five factors, accounting for 16.82, 11.71, 11.58, 11.39 and 8.56 per cent of the total variance respectively, a total of 60.05 per cent of the total variance explained, i.e. a high degree of total explained variance. The factor loadings are presented in Table XXX [give the table a number]. To enhance the interpretability of the factors, only variables with factor loadings as follows were selected for inclusion in their respective factors: >0.51 (factor 1), >0.44 (factor 2), >0.55 (factor 3), >0.44 (factor 4), and >0.64 (factor 5). The factors are named, respectively: (i) Leadership skills in school management; (ii) Parent and teacher partnerships in school development; (iii) Promoting staff development by creativity and consultation; (iv) Respect for, and confidence in, the senior management; and (v) Encouraging staff development through participation in decision making.

So far this reports only the process and the outcome of the data analysis. This would then have to be accompanied by a subsequent commentary on what the results *mean*, what they *show* and what are the *educational* aspects of the results: what the results show, suggest and what can be concluded from them. This would relate to, for example, the purposes of the research and the research questions, the main findings of the research, the theoretical and conceptual contributions of the results to the research, what can safely be concluded from the results and what alternative interpretations there are of the results. In other words, having presented the data for the factor analysis, the researcher then fits the results into the overall context and purposes of the research, interpreting and explaining the findings.

Box 43.1 provides the SPSS command sequence for conducting Principal Components Analysis.

43.2 What to look for in factor analysis output

SPSS typically produces many sets of data in factor analysis. The researcher must first conduct the 'safety checks', i.e. checking that the data are suitable for factor analysis, for example:

- sample size;
- number of variables;
- ratio of sample size to number of variables;
- interval and ratio data;
- normal distributions;
- linearity;
- outliers;
- selection bias/proper specification;
- theoretical underpinning of factors;
- intercorrelations between variables and intercorrelations between factors;
- strength of intercorrelations: most should be no less than 0.3 and no more than 0.6 (see the correlation matrix);

BOX 43.1 SPSS COMMAND SEQUENCE FOR PRINCIPAL COMPONENTS ANALYSIS

The SPSS command sequence for Principal Components Analysis is: 'Analyze' → 'Dimension Reduction' → 'Factor'. Send over to the 'Variables' box the variables which are to be included → Click the box marked 'Descriptives' (which opens another window) → In the area marked 'Correlation Matrix, check the boxes marked 'KMO and Bartlett's test of sphericity' and 'Coefficients' → Click 'Continue' → Click 'Extraction' (which opens another window) → Check that 'Principal components' is the setting in the area marked 'Method'. Ensure that the radio buttons are set for 'Correlation matrix' and 'Based on Eigenvalue' of 1. In the 'Display' check the boxes marked 'Unrotated factor solution' and 'Scree plot' → Click 'Continue' → Click 'Rotation' (which opens another window) → in the 'Method' area check either the 'Direct Oblimin' or 'Varimax' (depending on whether you wish to select the oblique rotation (Direct Oblimin) or the orthogonal rotation (Varimax)). In the 'Display' area, ensure that the 'Rotated Solution' box has been checked → Click 'Continue' → Click 'Options' (which opens a new window). In the 'Missing values' area, click the radio button 'Exclude cases pairwise', and in the 'Coefficient Display format' area check the box marked 'Sorted by size' → Click 'Continue' → Click 'OK'.

- Bartlett's test of sphericity should be statistically significant ($\rho < 0.05$);
- Kaiser-Mayer-Olkin measure of sampling adequacy should be 0.6 or higher.

If the safety checks indicate that it is safe to continue then what follows below is a set of pointers for what to look at in the different tables that SPSS typically produces. The example uses a Direct Oblimin rotation, taking an example of research on factors that affect teacher stress. In some cases the SPSS tables are too large to reproduce in their entirety, so extracts are included that illustrate the main points being made.

Imagine that we have given SPSS all the instructions indicated above, to run the SPSS analysis, and to check for the suitability of the data for factorization. In Table 43.3 the researcher checks that most of the correlation coefficients in the cells are greater than 0.3.

In Table 43.4 the researcher checks the suitability of the data for factor analysis by examining the output concerning the Kaiser-Meyer-Olkin (KMO) and the Bartlett test. Here the KMO measure is greater than 0.6 (0.845) and the Bartlett test is statistically significant (0.000), so the researcher is safe to continue, knowing that the data are suitable for factorization.

Table 43.5 indicates the amount of variance explained by each item (if it is lower than 0.3 then the item is a poor fit).

Table 43.6 indicates that two factors have been extracted (two components), i.e. those with Eigenvalues over 1: factor 1 explains 45.985 per cent of the total variance; factor 2 explains 18.852 per cent of the total

TABLE 43.3 CHECKING THE CORRELATION TABLE FOR SUITABILITY OF THE DATA FOR FACTORIZATION (SPSS OUTPUT)

		How much do you feel that working with colleagues all day is really a strain for you?	How much do you feel emotionally drained by your work?	How much do you worry that your job is hardening you emotionally?	How much frustration do you feel in your job?
Correlation	How much do you feel that working with colleagues all day is really a strain for you?	1.000	.554	.507	.461
	How much do you feel emotionally drained by your work?	.554	1.000	.580	.518
	How much do you worry that your job is hardening you emotionally?	.507	.580	1.000	.646
	How much frustration do you feel in your job?	.461	.518	.646	1.000

TABLE 43.4 CHECKING THE SUITABILITY OF THE DATA FOR FACTOR ANALYSIS (SPSS OUTPUT)

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.845
Bartlett's Test of Sphericity	Approx. Chi-Square	5460.475
	df	36
	Sig.	.000

variance. The total variance explained is 64.836 per cent (Cumulative percentage), i.e. very high.

If a scree plot has been produced then the researcher can look to see the 'bend in the elbow', i.e. the change in the shape of the plot, where the line flattens out and where the Eigenvalues are above and below 1.

Table 43.7 provides the pattern matrix, from which the researcher can identify which variables load onto the factors. The left-hand column prints the names of each variable and each column under the title 'Component' is a factor. Here we have emboldened and circled the variables that load onto each factor, for ease of identification.

Guidelines for which variables to select to include in each factor are:

- include the highest-scoring variables (those with the highest factor loadings);
- omit the low-scoring variables;
- look for where there is a clear scoring distance between those included and those excluded;
- review your selection to check that no lower-scoring variables have been excluded which are conceptually close to those included;
- review your selection to check whether some higher-scoring variables should be excluded if they are not sufficiently conceptually close to the others that have been included;
- review your final selection to see that the variables included in the factor are conceptually similar;
- each factor should include a minimum of three or four variables if possible (in the example in Table 43.7 it is clear that only two variables should be included).

Deciding on inclusions and exclusions is an art, not a science; there is no simple formula, so the researcher has to use his/her judgement.

In reporting the factor analysis the researcher should consider:

- reporting the method of factor analysis used (Principal Components; Direct Oblimin; KMO and Bartlett test of sphericity; Eigenvalues greater than 1; scree test; rotated solution);
- reporting how many factors were extracted with Eigenvalues greater than 1;
- reporting how many factors were included as a result of the scree test;
- giving a name/title to each of the factors;
- reporting how much of the total variance was explained by each factor;
- reporting the cut-off point for the variables included in each factor;
- reporting the factor loadings of each variable in the factor;
- reporting what the results show.

43.3 Cluster analysis

Whereas factor analysis enables the researcher to group *variables* into factors, cluster analysis enables the researcher to group together similar and homogeneous sub-samples of *people* (also termed 'cases'). In educational research this can be used, for example, to identify students with particular needs or abilities, regions with outstanding performance, high-achieving students, teachers with particular interests etc. Cluster analysis is often used in combination with factor analysis.

TABLE 43.5 CHECKING THE VARIANCE EXPLAINED BY EACH ITEM (SPSS OUTPUT)

Communalities

	Initial	Extraction
How hard do you feel you are working in your job?	1.000	.779
How much do you feel exhausted by the end of the workday?	1.000	.818
How much do you feel that you cannot cope with your job any longer?	1.000	.578
How much do you feel that you treat colleagues as impersonal objects?	1.000	.578
How much do you feel that working with colleagues all day is really a strain for you?	1.000	.602
How much do you feel emotionally drained by your work?	1.000	.629
How tired do you feel in the morning, having to face another school day?	1.000	.595
How much do you worry that your job is hardening you emotionally?	1.000	.661
How much frustration do you feel in your job?	1.000	.595

Extraction Method: Principal Component Analysis.

Cluster analysis is based on the measured proximity, distance and similarity between cases or variables. Distance/similarity, for example, is measured by Euclidean distance and Squared Euclidean distance (and SPSS constructs a proximity matrix from data), and clusters are formed by repeatedly combining the two or more cases with the greatest similarity. Cluster analysis has two main forms: hierarchical and non-hierarchical analysis (and, indeed, the two may be used in combination). Typically researchers can use K-means cluster

TABLE 43.6 EXTRACTION OF TWO FACTORS (SPSS OUTPUT)

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	4.139	45.985	45.985	4.139	45.985	45.985	4.028
2	1.697	18.851	64.836	1.697	18.851	64.836	1.991
3	.661	7.342	72.178				
4	.542	6.023	78.202				
5	.531	5.900	84.102				
6	.451	5.006	89.107				
7	.395	4.390	93.497				
8	.323	3.593	97.090				
9	.262	2.910	100.000				

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

TABLE 43.7 PATTERN MATRIX (SPSS OUTPUT WITH MARKINGS ADDED)

Pattern Matrix[a]

	Component	
	1	2
How hard do you feel you are working in your job?	.005	(.882)
How much do you feel exhausted by the end of the workday?	.252	(.834)
How much do you feel that you cannot cope with your job any longer?	(.691)	.234
How much do you feel that you treat colleagues as impersonal objects?	(.674)	-.459
How much do you feel that working with colleagues all day is really a strain for you?	(.782)	-.158
How much do you feel emotionally drained by your work?	(.774)	.096
How tired do you feel in the morning, having to face another school day?	(.697)	.247
How much do you worry that your job is hardening you emotionally?	(.814)	-.008
How much frustration do you feel in your job?	(.752)	.097

Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 6 iterations.

(non-hierarchical cluster analysis), hierarchical cluster and two-step cluster, of which hierarchical cluster analysis is the most widely used. Cluster analysis can work with interval, ratio, ordinal and nominal data, using different statistics for each kind of data. Here we comment on hierarchical cluster analysis, a commonly used method.

In hierarchical cluster analysis, homogeneous clusters of similar cases are formed at increasingly high levels of generality, and these are shown on a dendrogram. A dendrogram is a tree diagram that shows how cases are combined and linked at increasingly hierarchical and general levels until they become a single cluster, and it separates and indicates the several levels of combination. The dendrogram is an important feature of cluster analysis because it includes and lists all the cases, it indicates the level of similarity at which any two or more clusters are joined and the distance at which they are joined (the position of the line on the scale). Further, given its diagrammatic nature, it shows the clusters and their linkages at a glance and is easily understood.

Cluster analysis is best approached through software packages such as SPSS, and we illustrate this here. SPSS creates a dendrogram of results, grouping and regrouping groups until all the variables are embraced.

For example, here is a simple cluster analysis based on twenty cases (people). Imagine that their scores have been collected on an item concerning the variable 'your confidence in handling new situations'. Figure 43.3 presents the dendrogram of the clusters. Here one can see that, at the most general level there are two clusters: cluster one = persons 19, 20, 10, 11, 2, 4, 9, 3, 17, 18, 1, 16, 14, 15, 12, 13; cluster two = persons 7, 8, 5, 6. If one wished to have more detailed, specific clusters then three groupings can be found: cluster one: persons 19, 20, 10, 11, 2, 4, 9, 3, 17, 18, 1, 16; cluster two: persons 14, 15, 12, 13; cluster three: persons 7, 8, 5, 6. If one wishes to have a yet more specific, discriminating specification of groups then five groupings can be found: cluster one: persons 19, 20; cluster two: persons 10, 11, 2, 4, 9, 13; cluster three: persons 17, 18, 1, 16; cluster four: persons 14, 15, 12, 13; cluster five: persons 7, 8, 5, 6. In Figure 43.3, the final level in the hierarchy, bringing all the clusters into a single group, is not advised, not only because it defeats the purpose of identifying separate clusters but because the distance (shown in the horizontal axis as the distance from point

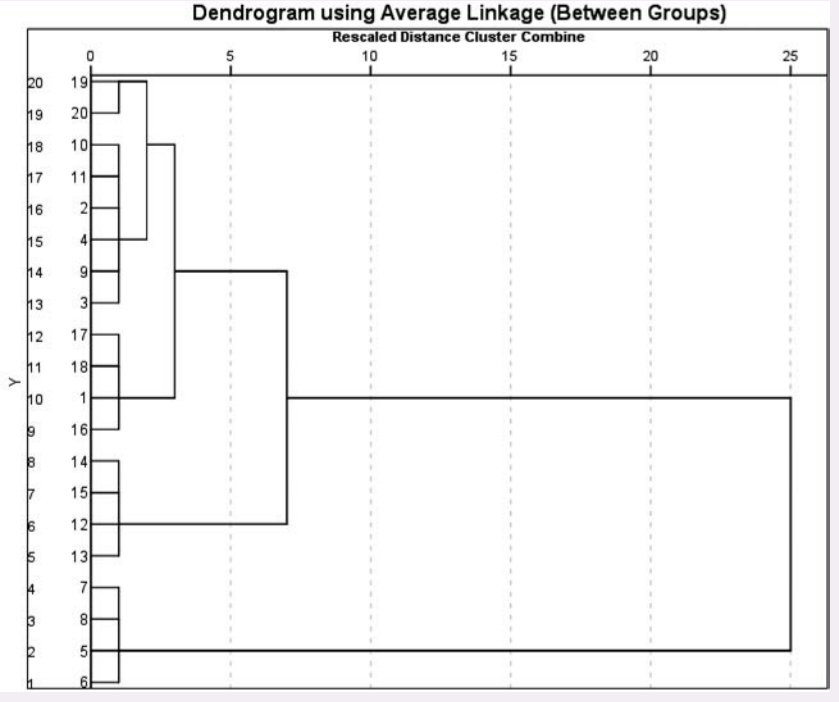


FIGURE 43.3 Cluster analysis using average linkage (SPSS output)

7 to point 25) is so great as to suggest gross dissimilarity, therefore the researcher here would be advised not to proceed beyond the two-cluster level (the second level) of the hierarchy, and in fact a strong case could be made for remaining at the first level of five clusters.

Before conducting cluster analysis the same 'safety checks' have to be performed as for factor analysis, with particular attention paid to the removal of outliers. If scales are different or if their standard deviations are large then standardized scores should be used.

Using cluster analysis enables the researcher to identify important groupings of people in a *post hoc* analysis, i.e. not setting up the groupings and sub-groupings at the stage of sample design, but *after* the data have been gathered. In the example of the two-group cluster here one could examine the characteristics of those participants who were clustered into groups one and two, and, for the three-group cluster, one could examine the characteristics of those participants who were clustered into groups one, two and three for the variable 'your confidence in handling new situations'. For example, in the two-group cluster it may be that group one comprises generally confident people whilst group two comprises generally less-confident people. For the three-group cluster it may be that one group comprises people below the age of twenty-one, the second group comprises people from ages twenty-two to forty, and the third group comprises people between the ages of forty-one and sixty-five.

There are several ways of conducting hierarchical cluster analysis; each way can produce a different clustering. For example, Figure 43.3 is the dendrogram that uses the default settings on SPSS, whilst Figure 43.4 uses a 'nearest neighbor' ('single linkage') setting.

As we can see, the two figures (43.3 and 43.4) are very different in the make-up of the groups (which people are in which groups), the linkages between groups and the number of 'layers' of groupings (three layers in Figure 43.3 and five layers in Figure 43.4). This means that researchers need to choose the most appropriate method of cluster analysis to use. This is beyond the scope of the present volume. Field (2000) and Yim and Ramdeen (2015) provide a straightforward introduction to key features of cluster analysis, showing clearly the steps to be used in SPSS, whilst Liew (2013) provides further introductory material on this. A useful introduction is provided at www.slideshare.net/jewelmrefran/cluster-analysis-15529464, with clear instructions for running cluster analysis with SPSS, and YouTube contains many short guides on working with cluster analysis.

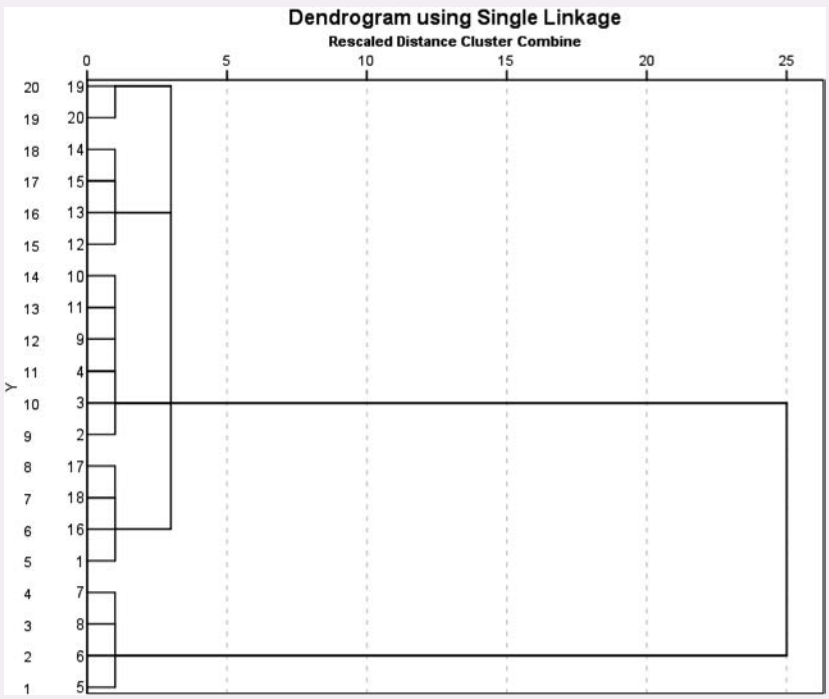


FIGURE 43.4 Cluster analysis using 'nearest neighbour' single linkage (SPSS output)

Box 43.2 provides the SPSS command sequence for hierarchical cluster analysis.

In reporting the results of cluster analysis, researchers must indicate the method of cluster analysis used and why, and what the results of the dendrogram show. It is important for researchers to indicate:

- what is the similarity criterion that combines individual cases into a single cluster (e.g. all females);
- how many cases (and who) are in each cluster;
- how similar the cases are within each cluster. The proximity matrix in SPSS indicates the measured degree of closeness or distance between cases, and this is useful in enabling the researcher to comment on just how similar the cases are within the cluster;
- what differentiates that cluster from another (e.g. high achievers in one cluster and low achievers in another), and, as clusters are combined further up the hierarchy of the dendrogram, what is the criterion or characteristic that combines them (this is based on the judgement of the researcher rather than a statistical procedure);
- how similar/dissimilar are the clusters. Again, the proximity matrix in SPSS indicates the measured degree of closeness or distance between clusters, and this is useful in enabling the researcher to comment on how different one cluster is from another;
- what is the most suitable 'cut-off' point in a dendrogram and why, i.e. at what level in the hierarchy it is most advisable to cease combining clusters. This requires the researcher to make an educational rather than statistical judgement, though the distance criterion (shown on the horizontal distance on an SPSS dendrogram) is a useful guide here.

Cluster analysis has been criticized for being atheoretical, non-generalizable, descriptive rather than inferential, its assumption that real structures exist (i.e. clusters are based on statistical rather than actual similarity), and for being more suited to small rather than large samples (the latter being particularly a problem in constructing a dendrogram). Nevertheless it is a frequently used technique for grouping cases, and working with hierarchical cluster analysis in SPSS is a useful tool for researchers. For further guidance we refer readers to Field (2000), Landau and Chis Ster (2010); Everitt *et al*. (2011), Liew (2013) and Yim and Ramdeen (2015).

43.4 A note on structural equation modelling

Structural equation modelling (SEM) is a further method in statistics-based research using interval and ratio data. We provide an introduction to it here; for further in-depth discussion researchers should go sources referenced below. Structural equation modelling is the name given to a group of techniques that enable researchers to construct models of putative causal relations, and to test those models against data. It is designed to enable researchers to confirm, modify and test their models of causal relations between variables. It is based on multiple regression, but advances beyond this to create and test models of relationships, often causal, to see how well the models fit the data, and in this respect it is often also used for confirmatory factor analysis. Though SPSS does not have a function to handle this, the Analysis of Moment Structures (AMOS) is a software package that enables the researcher to import and work with SPSS files.

As mentioned earlier, factor analysis can be both *exploratory* and *confirmatory*. Structural equation modelling is used in confirmatory factor analysis. Whilst the earlier discussion concerned exploratory factor analysis, confirmatory factor analysis is a feature of the group of latent variable models (models of factors

BOX 43.2 SPSS COMMAND SEQUENCE FOR HIERARCHICAL CLUSTER ANALYSIS

The SPSS command sequence for hierarchical cluster analysis is: 'Analyze' → 'Classify' → 'Hierarchical cluster' → Send over the variable(s) of interest to the box marked 'Variable(s)' → Click the box marked 'Statistics', which this opens a new window, and check the box marked 'Proximity matrix' → Click 'Continue' → Click the box marked 'Plots', which opens a new window, and check the box marked 'Dendrogram' → Click 'Continue' → Click the box marked 'Method', which opens a new window, and decide, from the drop-down menus, which 'Cluster Method' you wish to use (the default setting is 'Between groups linkage', which is widely used, but many researchers also suggest using the advocate the 'Nearest neighbor' setting. Decide which 'Measure' you wish to use (for scale data – the 'Interval' radio button – a widely used measure is the default setting of 'Square Euclidian Distance'; for ordinal data (the ('Counts' radio button) the chi-square setting is often used; and for binary nominal data (the 'Binary' radio button) the 'Squared Euclidian Distance' is often used) → Click 'Continue' → Click 'OK'.

rather than observed variables) which includes factor analysis, path analysis and structural equation analysis. Confirmatory factor analysis seeks to verify (to confirm or refute) the researcher's predictions about factors and their factor loadings in data and data structures. As mentioned earlier, factors are latent, they cannot be observed as they underlie variables.

By contrast, path analysis – an extension of multiple regression – only works with observed variables, and it attempts to estimate and test the magnitude and significance of relationships, often putatively causal, between sets of observed variables. Path analysis is a statistical method that enables a researcher to determine how well a multivariate set of data fits with a particular (causal) model that has been set up in advance by the researcher (i.e. an *a priori* model). It is a particular kind of multiple regression analysis that enables the researcher to see the relative weightings of observed independent variables on each other and on a dependent variable, to establish pathways of causation, and to determine the direct and indirect effects of independent variables on a dependent variable (Morrison, 2009, p. 96). The researcher constructs what she or he thinks will be a suitable model of the causal pathway between independent variables and between independent and dependent variables, often based on literature and theory, and then tests this to see how well it fits with the data.

In constructing path analysis, computer software is virtually essential. Programs such as AMOS (in SPSS) and LISREL are two commonly used examples.

Morrison (2009, pp. 96–8) gives an example of path analysis in degree classification, with three independent variables and their relationship to the dependent variable of degree classification:

- socio-economic status
- part-time working
- level of motivation for academic study

These three variables are purported to have an effect on the class of degree that a student gains (the dependent variable). Figure 43.5 is constructed from the AMOS software.

The researcher believes that this can be modelled in a nonrecursive model (a model in which the direction of causality is not solely one-way – see the direction of the causal arrows joining 'part-time work' and 'level of motivation for academic study' in Figure 43.5, which go to and from each other. This is in contrast to a recursive model, in which the direction of putative causality is one-way only). In the nonrecursive model here, socio-economic status determines part-time working, level of motivation for academic study and the dependent variable 'class of degree'. The variable 'socio-economic status' is deemed to be an exogenous variable (a variable caused by variables that are *not* included in the causal model), whilst the variables 'part-time working' and 'level of motivation for academic study' are deemed to be endogenous variables (those caused by variables that *are* included in the model) as well as being affected by exogenous variables.

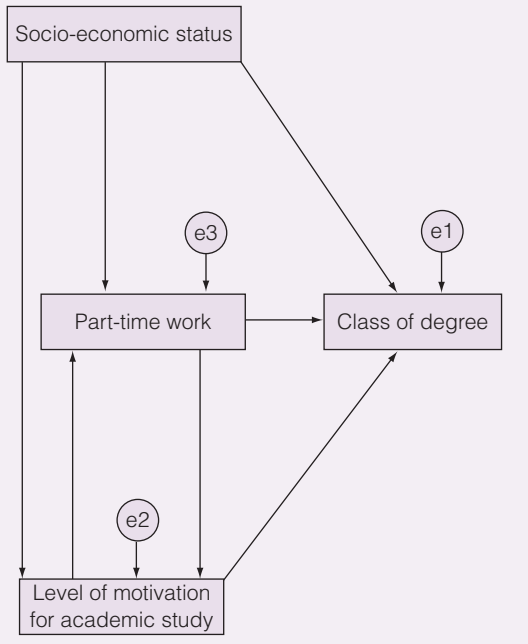


FIGURE 43.5 Path analysis modelling with AMOS (AMOS output)

In the model (Figure 43.5), the dependent variable is 'class of degree' and there are directional causal arrows leading both to this dependent variable and to and from the three independent variables. The model assumes that the variables 'part-time work' and 'level of motivation for academic study' influence each other and that socio-economic status precedes the other independent variables rather than being caused by them. In the model there are also three variables in circles, termed 'e1', 'e2' and 'e3'; these three additional variables are 'error factors', i.e. additional extraneous/exogenous factors which may also be influencing the three variables in question, and AMOS adjusts the results for these factors. (AMOS also enables the researcher to draw the model and manipulate its layout.)

AMOS then calculates the regression coefficient of each relationship and places each coefficient on the model. An example of the model generated by AMOS

is presented in Figure 43.6 (using the 'standardized estimates' in AMOS).

From Figure 43.6 one can see that:

- 'socio-economic' status exerts a direct influence on class of degree (0.18), and that this is higher than the direct influence of either 'part-time work' (–.01) or 'level of motivation for academic study' (0.04);
- 'socio-economic status' exerts a powerful direct influence on 'level of motivation for academic study' (0.52), and this is higher than the influence of 'socio-economic status' on 'class of degree' (0.18);
- 'socio-economic status' exerts a direct and negative influence on 'part-time work' (–0.21), i.e. the higher the socio-economic status, the lesser is the amount of part-time work undertaken;
- 'part-time work' exerts a powerful direct influence on 'level of motivation for academic study' (1.37), and this is higher than the influence of 'socio-economic status' on 'level of motivation for academic study' (0.52);
- 'level of motivation for academic study' exerts a powerful negative direct influence on 'part-time work' (–1.45), i.e. the higher is the level of motivation for academic study, the lesser is the amount of part-time work undertaken;

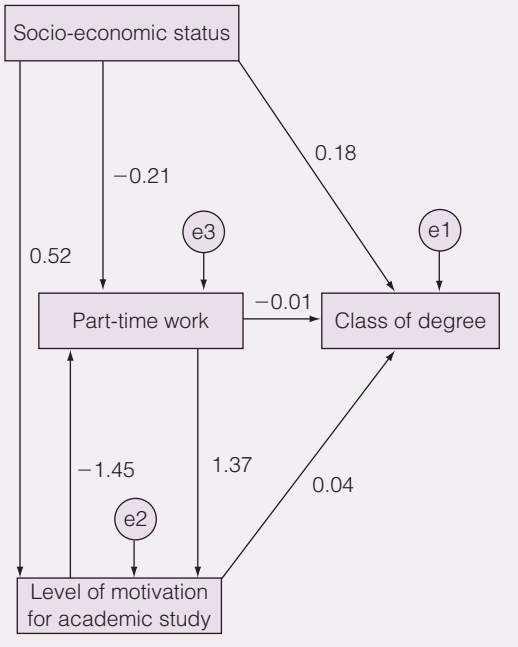


FIGURE 43.6 Path analysis with calculations added (AMOS output)

- 'level of motivation for academic study' exerts slightly more influence on 'class of degree' (0.04) than does 'part-time work' (–0.01);
- 'part-time work' exerts a negative influence on the class of degree (–0.01), i.e. the more one works part-time, the lower is the class of degree obtained.

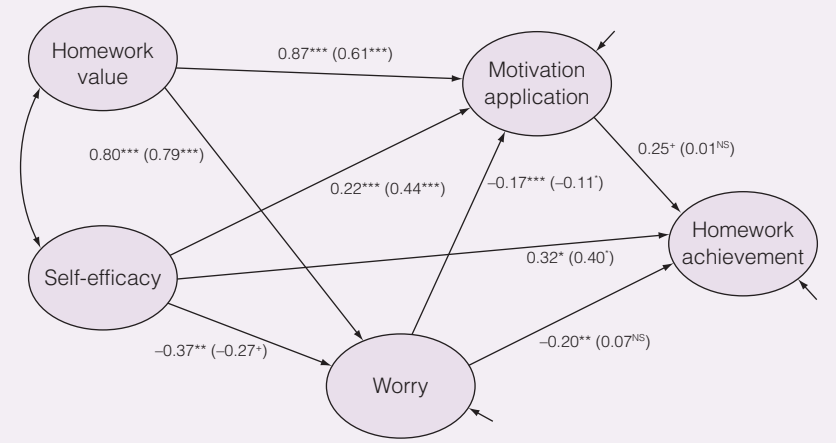
AMOS also yields a battery of statistics about the 'goodness of fit' of the model to the data, most of which are beyond the scope of this book. Kline (2015a, pp. 268–80) suggests three main tests of goodness of fit here: the chi-square statistic, the Comparative Fit Index (CFI) and the Root Mean Square Error of Approximation (RMSEA). Suffice it to say here that the chi-square statistic must *not* be statistically significant (i.e. $\rho > 0.05$), i.e. to indicate that the model does not differ statistically significantly from the data, in other words that the model is faithful to the data, the Comparative Fit Index (CFI) should be 0.9 or higher and the Root Mean Square Error of Approximation (RMSEA) should be below 0.10, and ideally should be smaller than 0.05.

Hong *et al*. (2015) provide a clear example of a structural equation model of the causal effects of homework motivation and worry anxiety on homework achievement in mathematics and English (Figure 43.7). The model in Figure 43.7 derived from theories of social-cognitive and expectancy-value theories of motivation, and included antecedent variables (perceived homework value and self-efficacy), mediating effects and direct effects (worry and motivation application).

The authors constructed a causal model, then tested its goodness of fit and then considered the findings. They report an acceptable goodness of fit: for mathematics the chi-square was not statistically significant ($\rho = 0.011$), the CFI was 0.997 and the RMSEA was 0.048; for English the chi-square was not such a good fit, being statistically significant ($\rho = 0.0005$), whereas the CFI was 0.990 and the RMSEA was 0.073, both of these being a good fit.

Having conducted a goodness of fit test, the authors then report the findings. In Figure 43.7 the figures outside the brackets are the beta weightings for mathematics and those inside the brackets are for English. Here the strongest direct effect on homework achievement was from self-efficacy (0.32 for mathematics and 0.40 for English) and the antecedent variable of perceived homework value had a very strong effect on motivation application (0.87 for mathematics and 0.61 for English) and on worry (0.80 for mathematics and 0.79 for English).

Path analysis assumes that the *direction* of causation in the variables can be identified, that the data are at the interval or ratio level, that the relations are linear, that



The final structural model with standardized path coefficients.
$^{+}p < 0.10$. $^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$. NS = Not significant

FIGURE 43.7 A structural equation model of homework motivation and worry on homework achievement

Source: Hong *et al.* (2015, p. 496)

Note

The figures outside the brackets refer to mathematics and those inside the brackets refer to English

the data meet the usual criteria for regression analysis, and that the model's parsimony (inclusion of few variables) is fair. Morrison (2009, p. 98) argues that path analysis is only as good as the causal assumptions that underpin it, nor does it prove unequivocally that causation is present; rather it only tests a model based on *assumed* causal directions and influences. Nevertheless, its utility lies in its ability to test models of putative causal directions, to establish relative weightings of variables, to look at direct and indirect effects of independent variables and to handle several independent variables simultaneously.

Structural equation modelling combines the features of confirmatory factor analysis (i.e. it works with *latent* factors) and of path analysis (i.e. it works with *observed*, manifest variables). Here each factor is a latent construct comprising several variables. This is shown in Figure 43.8, in which each factor appears in the ovals and each observed variable appears in a rectangle. Here the factor 'socio-economic status' (S) has three variables (S1, S2, S3), the factor 'part-time work' (P) has two variables (P1, P2) and the factor 'level of motivation for academic study' (L) has three variables (L1, L2, L3). Each variable has its own error factor (the small circles with 'E' inside them).

Structural equation modelling requires the researcher to:

- construct the model (the factors and the variables);
- decide the direction of causality (recursive or non-recursive);
- identify the number of parameters to be estimated (number of factor coefficients, covariances, observations);
- run the AMOS analysis (or another piece of software) and check the goodness of fit of the model to the data;
- make any necessary modifications to the model;
- report the findings.

This is a highly simplified overview of the process and nature of structural equation modelling, and the reader is strongly advised to read further, more details texts, e.g. Loehlin (2004), Schumacker and Lomax (2004), Kline (2005a; 2015) and Tabachnick and Fidell (2013).

43.5 A note on multilevel modelling

Multilevel modelling (also known as multilevel regression and hierarchical modelling) recognizes that individual characteristics are nested within group characteristics and, indeed, wider contextual factors and that these can be factored into data analysis simultaneously. Students are nested within classrooms, classrooms are nested within schools, schools are nested within neighbourhoods and so on. There are many

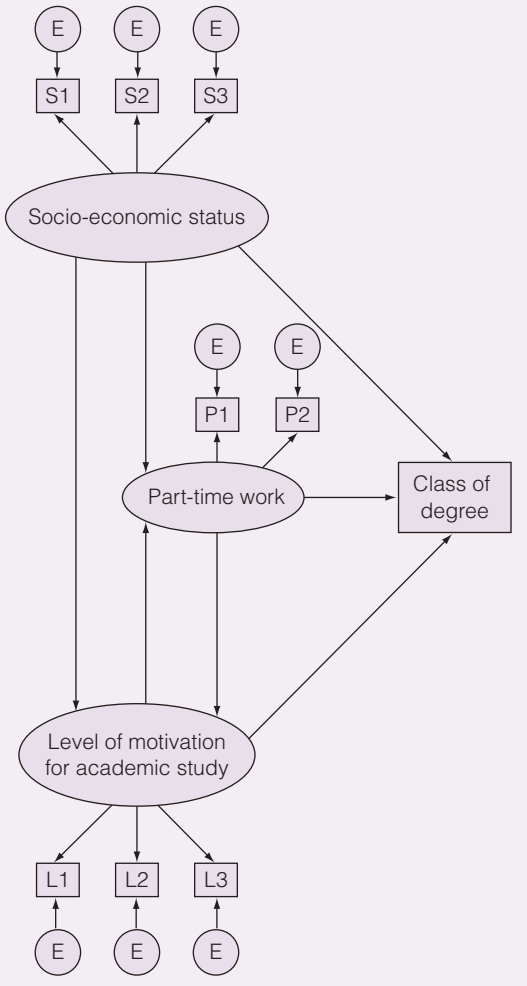


FIGURE 43.8 A structural equation model

levels of influences on outcomes or observed findings. Multilevel analysis addresses this (cf. Bickel, 2007; Seltzer and Rickles, 2012; Robson and Pevalin, 2016), recognizing that, if multilevel nesting is not addressed in statistical analysis then there is a risk of finding spurious results (e.g. statistically significant differences between individuals or groups) or of overlooking differences between smaller groups (e.g. minority groups). Multilevel analysis is frequently used in school effects and effectiveness research,

Typically in most schools, students are brought together in particular groupings for specified purposes and each group of students has its own different characteristics which render it different from other groups. Multilevel modelling addresses the fact that, unless it can be shown that different groups of students are, in fact, alike, it is generally inappropriate to aggregate groups of students or data for the purposes of analysis. Multilevel models avoid the pitfalls of aggregation and the *ecological fallacy* (Plewis, 1997, p. 35), i.e. making inferences about individual students and behaviour from aggregated data.

Data and variables exist at individual and group levels, indeed Keeves and Sellin (1997b) break down analysis further into three main levels: (a) between students over all groups; (b) between groups; and (c) between students within groups. One can extend the identification of levels, of course, to include individual, group, class, school, local, regional, national and international levels (Paterson and Goldstein, 1991). Data are 'nested' (Bickel, 2007), i.e. individual-level data are nested within group, class, school, regional etc. levels; a dependent variable is affected by independent variables at different levels (p. 3). In other words, data are hierarchical. If we are looking at, say, the effectiveness of a reading program in a region, we must recognize that student performance at the individual level is also affected by group and school level factors (e.g. differences within a school may be smaller than differences between schools). Individuals within families may be more similar than individuals between families. Using multilevel modelling researchers can ascertain, for example, how much of the variation in student attainment might be attributable to differences within students in a single school or to differences between schools (i.e. how much influence is exerted on student attainment by the school which the student attends). Another example might be the extent to which factors such as sex, ethnicity, type of school, locality of school, school size account for variation in student performance. Multilevel modelling enables the researcher to calculate the relative impact on a dependent variable of one or more independent variables at each level of the hierarchy, and, thereby to identify factors at each level of the hierarchy that are associated with the impact of that level.

Multilevel modelling has been conducted using multilevel regression and hierarchical linear modelling (HLM). It enables researchers to ask questions hitherto unanswered, e.g. about variability between and within schools, teachers and curricula, in short about the *processes* of teaching and learning. Multilevel analysis avoids statistical treatments associated with experimental methods (e.g. analysis of variance and covariance); rather it uses regression analysis and, in particular, multilevel regression. Regression analysis assumes homoscedasticity (where the residuals demonstrate equal scatter), that the residuals are independent of each

other, and finally, that the residuals are normally distributed.

Multilevel modelling is the basis of much research on the 'value added' component of education and the comparison of schools in public 'league tables' of results. It is not without its critics, e.g. Gorard (2007, p. 221) argues that multilevel modelling has 'an unclear theoretical and empirical basis', is unnecessarily complex, that it has not produced any important practical research results, and that, due to the presence of alternatives, is largely unnecessary, with limited ease of readability by different audiences. Nonetheless, multilevel modelling attracts worldwide interest.

Whereas ordinary regression models do not make allowances, for example, for different schools (Paterson and Goldstein, 1991), multilevel regression can include school differences, and indeed other variables, e.g.: socio-economic status, single and co-educational schools, location, size of school, teaching styles etc. Indeed Bickel (2007) and Seltzer and Rickles (2012) indicate how multilevel modelling can be used in longitudinal studies.

The Bristol Centre for Multilevel Modelling (www.bristol.ac.uk/cmm/) produces online courses, introductory materials, workshops and downloads for multilevel modelling, software downloads for conducting multilevel modelling, and full sets of references and papers. Further materials and references can be found at Scientific Software International.

Useful overviews of multilevel modelling can be found in Goldstein (1987, 2003), Raudenbush and Bryk (2002), Keeves and Sellin (1997b), Bickel (2007), O'Connell and McCoach (2008), Snijders and Bosker (2012), Tabachnick and Fidell (2013), and Robson and Pevalin (2016), whilst Heck, Thomas and Tabaska (2013) provide a comprehensive, if demanding, introduction to multilevel modelling with SPSS. There are useful publications from the Bristol Centre for Multilevel Modelling (www.bristol.ac.uk/cmm/research/).

Companion Website

The companion website to the book provides additional material, data files and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Choosing a statistical test

CHAPTER 44

Having set out a range of statistical tests in the preceding chapters, this chapter guides the researcher on which tests to use with particular kinds of data and for specific purposes, i.e. to address fitness for purpose. The chapter proceeds thus, indicating different considerations which researchers must bear in mind when selecting the most appropriate tests:

- how many samples?
- the types of data used
- choosing the right statistic
- assumptions of tests

The chapter provides several tables to guide the researcher in making choices here.

44.1 Introduction

There are very many statistical tests available to the researcher. Which test one employs depends on several factors, for example:

- the purpose of the analysis (e.g. to describe and/or explore data, to test a hypothesis, to seek correlations, to identify the effects of one or more independent variables on a dependent variable, to identify differences between two or more groups, to look for underlying groupings of data, to report effect sizes);
- the kinds of data with which one is working (parametric and non-parametric);
- the scales of data being used (nominal, ordinal, interval, ratio);
- the number of groups in the sample;
- the assumptions in the statistical tests;
- whether the samples are independent of each other or related to each other.

Researchers wishing to use statistics will need to ask and answer questions such as:

- What statistics do I need to answer my research questions?
- What kind of statistic do I need (e.g. difference test, correlation, factor analysis, regression, grouping (of variables, of people);
- Are the data parametric or non-parametric?
- How many groups are there (e.g. two, three or more)?
- Are the groups related or independent?
- Have all the assumptions of the statistical test been met?

We have addressed these points in the preceding chapters. In this chapter we draw together the threads of the discussion of statistical analysis and address what, for many researchers, can be challenging: deciding which statistical tests to use. In the interests of clarity we use tables and graphics for presenting the issues in this chapter.

44.2 Sampling issues

In addition to the scale of data being used (nominal, ordinal, interval, ratio), the *kind* of statistic that one calculates depends in part on: first, whether the samples are related to, or independent of, each other; second, the number of samples in the test; and third, whether the assumptions (the 'safety checks' set out in previous chapters) have been met. With regard to the first point, as we have seen in previous chapters, different statistics are sometimes used when groups are related to each other and when they are independent of each other. Groups are independent when they have no relationship to each other, for example, in conducting a test to see if there is any difference between the voting of males and females on a particular item, say mathematics performance. The tests that one could use here are, for example: the chi-square test (for nominal data), the Mann-Whitney U test and Kruskal-Wallis (for ordinal data), and the t-test and Analysis of Variance (ANOVA) for interval and ratio data.

However, there are times when the groups might be related. For example, we may wish to measure the performance of the same group at two points in time – before and after a particular intervention – or we may

wish to measure the voting of the same group on two different variables, say preference for mathematics and preference for music, or under two conditions (e.g. working in a noisy and a silent environment). Here it is not different groups that are being involved, but the same group on two occasions, two variables or two conditions respectively. In this case different statistics are used, for example, the Wilcoxon test, the Friedman test, the t-test for paired samples and the sign test. Table 44.1 gives a frequently used example of an experiment.

In preceding chapters we have indicated which tests are used with independent samples and which are used with related samples.

With regard to the number of samples/groups in the test, there are statistical tests which are for single samples (one group only, e.g. a single class in school), for two samples (two groups, e.g. males and females in a school) and for more than two samples, for example, parents, teachers, students and administrative staff in a school. Tests which can be applied to a *single* group include the binomial test, the chi-square one-sample test and the Kolmogorov-Smirnov one-sample test; tests which can be applied to *two* groups include the chi-square test, the Mann-Whitney U test, the t-test, the Spearman and Pearson tests of correlation; tests which can be applied to *three or more* samples include the chi-square test, ANOVA, the Kruskal-Wallis test and the Tukey and Games-Howell tests. We set out some of these tests in Table 44.2. It is essential to use the correct test for the correct number of groups.

TABLE 44.1 IDENTIFYING STATISTICAL TESTS FOR AN EXPERIMENT

CONTROL GROUP ↑ t-test for independent samples for the pre-test ↓ EXPERIMENTAL GROUP	Wilcoxon test or t-test for paired samples (depending on data type) Wilcoxon test or t-test for paired samples (depending on data type)	CONTROL GROUP ↑ t-test for independent samples for the post-test ↓ EXPERIMENTAL GROUP

TABLE 44.2 STATISTICAL TESTS TO BE USED WITH DIFFERENT NUMBERS OF GROUPS OF SAMPLES

Scale of data	*One sample*	*Two samples*		*More than two samples*	
		Independent	*Related*	*Independent*	*Related*
Nominal	Binomial	Fisher exact test	McNemar	Chi-square (χ^2) k-samples test	Cochran Q
	Chi-square (χ^2) one-sample test	Chi-square (χ^2) two-samples test			
Ordinal	Kolmogorov-Smirnov one-sample test	Mann-Whitney U test	Wilcoxon matched pairs test	Kruskal-Wallis test	Friedman test
		Kolmogorov-Smirnov test	Sign test	Ordinal regression analysis	
		Wald-Wolfowitz			
		Spearman rho			
		Ordinal regression analysis			
Interval and ratio	t-test	t-test	t-test for paired samples	One-way ANOVA	Repeated measures ANOVA
		Pearson product moment correlation		Two-way ANOVA	
				Tukey hsd test	
				Scheffé test	

44.3 The types of data used

The statistical tests used also depend on the scales of data being treated (nominal to ratio) and the tasks which the researcher wishes to perform – the purpose of the analysis (e.g. to discover differences between groups, to look for degrees of association, to measure the effect of one or more independent variables on a dependent variable etc.). In preceding chapters we have described the different scales of data and the kinds of tests available for different purposes. With these considerations, Table 44.3 summarizes some of the main tests here.

The type of test used also varies according to whether one is working with parametric or non-parametric data.

44.4 Choosing the right statistic

Figure 44.1 and Table 44.4 draw together and present the kinds of statistical tests available, depending on whether one is using parametric or non-parametric data, together with the purpose of the analysis. Table 44.4 sets out the commonly used statistics for data types and purposes.

44.5 Assumptions of tests

Statistical tests are based on certain assumptions. It is important to be aware of these assumptions and to operate fairly within them. Table 44.5 sets out the assumptions which need to be met (the 'safety checks' of previous chapters) if the statistics in question are to be used. Unfortunately researchers often use statistics for parametric tests (e.g. t-tests, ANOVA, regression) when the assumptions have not been met, and this undermines the results. If the assumptions underpinning a parametric test have not been met then it is often wiser to revert to an equivalent non-parametric test. Some of the more widely used tests have the assumptions set out (Table 44.5).

The choice of which statistics to employ is not arbitrary, and we have set out in this chapter the considerations that must be addressed in selecting the correct statistic.

TABLE 44.3 TYPES OF STATISTICAL TESTS FOR FOUR SCALES OF DATA

	Nominal	*Ordinal*	*Interval and ratio*
Measures of association	Tetrachoric correlation	Spearman's rho	Pearson product-moment correlation
	Point biserial correlation	Kendall rank order correlation	
	Phi coefficient	Kendall partial rank correlation	
	Cramer's *V*		
Measures of difference	Chi-square	Mann-Whitney U test	t-test for two independent samples
	McNemar	Kruskal-Wallis	t-test for two related samples
	Cochran Q	Wilcoxon matched pairs	One-way ANOVA
	Binomial test	Friedman two-way analysis of variance	Two-way ANOVA for more
		Wald-Wolfowitz test	Tukey hsd test
		Kolmogorov-Smirnov test	Scheffé test
Measures of linear relationship between independent and dependent variables		Ordinal regression analysis	Linear regression
Identifying underlying factors, data reduction, grouping of people or variables			Multiple regression Factor analysis Cluster analysis

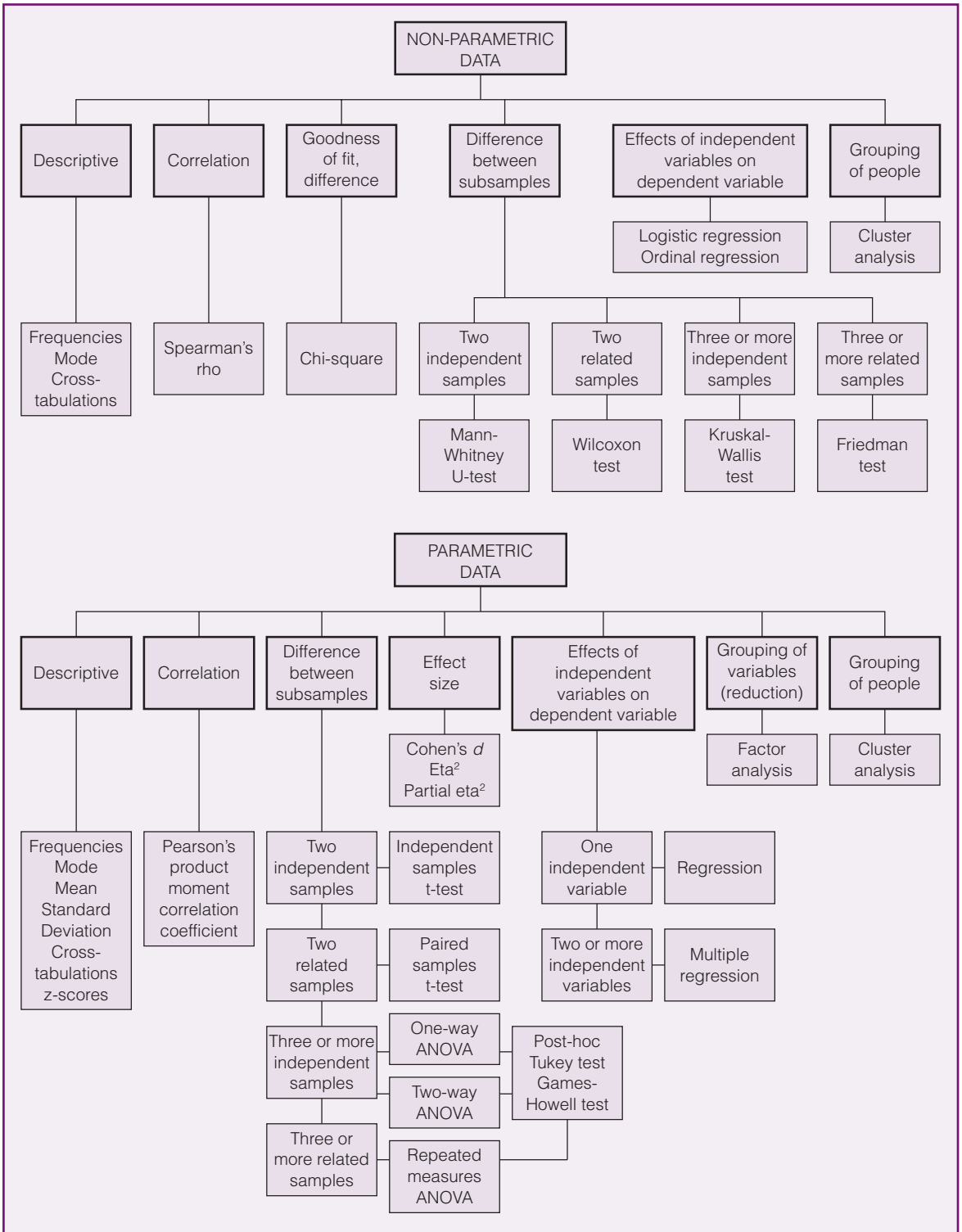


FIGURE 44.1 Choosing statistical tests for parametric and non-parametric data

TABLE 38.4 STATISTICS AVAILABLE FOR DIFFERENT TYPES OF DATA

Data type	*Legitimate statistics*	*Points to observe/questions/examples*
Nominal	i Mode (the score achieved by the greatest number of people)	Is there a clear 'front runner' that receives the highest score with low scoring on other categories, or is the modal score only narrowly leading the other categories? Are there two scores which are vying for the highest score – a bi-modal score?
	ii Frequencies	Which are the highest/lowest frequencies? Is the distribution even across categories?
	iii Chi-square (χ^2) (a statistic that charts the difference between statistically expected and actual scores)	Are differences between scores caused by chance/accident or are they statistically significant, i.e. not simply caused by chance?
Ordinal	i Mode	Which score on a rating scale is the most frequent?
	ii Median (the score gained by the middle person in a ranked group of people or, if there is an even number of cases, the score which is midway between the highest score obtained in the lower half of the cases and the lowest score obtained in the higher half of the cases).	What is the score of the middle person in a list of scores?
	iii Frequencies	Do responses tend to cluster around one or two categories of a rating scale? Are the responses skewed towards one end of a rating scale (e.g. 'strongly agree')? Do the responses pattern themselves consistently across the sample? Are the frequencies generally high or generally low (i.e. whether respondents tend to feel strongly about an issue)? Is there a clustering of responses around the central categories of a rating scale (the central tendency, respondents not wishing to appear to be too extreme)?
	iv Chi-square (χ^2)	Are the frequencies of one set of nominal variables (e.g. sex) significantly related to a set of ordinal variables?
	v Spearman rank order correlation (a statistic to measure the degree of association between two ordinal variables)	Do the results from one rating scale correlate with the results from another rating scale? Do the rank order positions for one variable correlate with the rank order positions for another variable?
	vi Mann-Whitney U-test (a statistic to measure any significant difference between two independent samples)	Is there a significant difference in the results of a rating scale for two independent samples (e.g. males and females)?
	vii Kruskal-Wallis analysis of variance (a statistic to measure any significant differences between three or more independent samples)	Is there a significant difference between three or more nominal variables (e.g. membership of political parties) and the results of a rating scale?

continued

TABLE 44.4 CONTINUED

Data type	*Legitimate statistics*	*Points to observe/questions/examples*
Interval and ratio	i Mode ii Mean iii Frequencies iv Median v Chi-square (χ^2)	What is the average score for this group?
	vi Standard deviation (a measure of the dispersal of scores)	Are the scores on a parametric test evenly distributed? Do scores cluster closely around the mean? Are scores widely spread around the mean? Are scores dispersed evenly? Are one or two extreme scores ('outliers') exerting a disproportionate influence on what are otherwise closely clustered scores?
	vii z-scores (a statistic to convert scores from different scales, i.e. with different means and standard deviations, to a common scale, i.e. with the same mean and standard deviation, enabling different scores to be compared fairly)	How do the scores obtained by students on a test which was marked out of 20 compare to the scores by the same students on a test which was marked out of 50?
	viii Pearson product moment correlation (a statistic to measure the degree of association between two interval or ratio variables)	Is there a correlation between one set of interval data (e.g. test scores for one examination) and another set of interval data (e.g. test scores on another examination)?
	ix t-tests (a statistic to measure the difference between the means of one sample on two separate occasions or between two samples on one occasion)	Are the control and experimental groups matched in their mean scores on a parametric test? Is there a significant difference between the pre-test and post-test scores of a sample group?
	x Analysis of variance (a statistic to ascertain whether two or more means differ significantly)	Are the differences in the means between test results of three groups statistically significant?
	xi Regression	What are the predicted scores on one variable if we know the scores on another variable?
	xii Multiple regression	What are the relative weightings of two or more independent variables on a dependent variable?
	xiii Factor analysis	What are the underlying, latent factors into which variables can be grouped?
	xiv Structural equation modelling	What is the causal model of relations between independent variables and factors on a dependent variable?

TABLE 44.5 ASSUMPTIONS OF STATISTICAL TESTS

Test	*Assumptions*
Mean	Data are normally distributed, with no outliers
Mode	There are few values, and few scores, occurring which have a similar frequency
Median	There are many ordinal values
Chi-square	Data are categorical (nominal); Randomly sampled population; Independent categories; Data are discrete (i.e. no decimal places between data points); 80% of all the cells in a crosstabulation contain 5 or more cases;
Kolmogorov-Smirnov	The underlying distribution is continuous; Data are nominal;
t-test and Analysis of Variance	Population is normally distributed; Sample is selected randomly from the population; Parametric data; Each group is independent of the other; The groups to be compared are nominal, and the comparison is made using interval and ratio data; The sets of data to be compared are normally distributed (the bell-shaped Gaussian curve of distribution); The sets of scores have approximately equal variances, or the square of the standard deviation is known; The data are interval or ratio.
Multiple Analysis of Variance (MANOVA)	Continuous parametric data for dependent variables; Independent variables are categorical, with two or more values; Independent groups; Random sampling; Adequate sample size (more cases in each cell than the number of dependent variables being studied); Normal distribution of the data; No outliers; Linear relationship between each pair of dependent variables; No multicollinearity (dependent variables are independent of each other but moderately correlated); Homogeneity (equality) of variances.
Wilcoxon Test	The data are ordinal; The samples are related.
Mann-Whitney and Kruskal-Wallis	The groups to be compared are nominal, and the comparison is made using ordinal data; The populations from which the samples are drawn have similar distributions; Samples are drawn randomly; Samples are independent of each other;
Spearman rank order correlation	The data are ordinal;
Pearson correlation	The data are interval and ratio;

continued

TABLE 44.5 CONTINUED

Test	*Assumptions*
Regression (simple and multiple)	The data derive from a random or probability sample; Adequate sample size; The data are interval or ratio (unless ordinal regression is used); Avoidance of singularity (where one variable is a combination of independent variables); Outliers have been removed; There is a linear relationship between the independent and dependent variables; The dependent variable is normally distributed (the bell-shaped Gaussian curve of distribution); The residuals for the dependent variable (the differences between calculated and observed scores) are approximately normally and consistently evenly distributed (homoscedasticity); Collinearity is removed (where one independent variable is an exact or very close correlate of another); The residuals are not strongly correlated with the independent variables; Each case is independent of the others.
Factor analysis	The data are interval or ratio; The data are normally distributed; Outliers have been removed; The sample size should not be less than 100–150 persons; There should be at least five cases for each variable; The relationships between the variables should be linear; *Intercorrelations between variables* should be between.3 and 0.6; The factors should not be highly correlated with each other (the principle of orthogonality); Each factor contains more than two variables; Strong theoretical underpinning of each factor; The data must be capable of being factored.
Cluster analysis	Same as for factor analysis except that data can be nominal ordinal, interval or ratio; If scales are different or if their standard deviations are large then standardized scores should be used.

Companion Website

The companion website to the book provides additional material, data files and PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Beyond mixed methods

CHAPTER 45

Using Qualitative Comparative Analysis (QCA) to integrate cross-case and within-case analyses

Barry Cooper and Judith Glaesser

This chapter provides a brief overview of situations in which methods might fruitfully be mixed before concentrating on one approach, Ragin's Qualitative Comparative Analysis (QCA). QCA facilitates the integration of quantitative and qualitative forms of analysis, providing a case-based method for selecting cases for in-depth analysis to supplement its rigorous cross-case analysis. QCA is based on set theory and Boolean algebra, relying on 'truth tables' as the basis of analysis. These truth tables provide a representation of data that enables researchers to focus easily on types of cases. The foundations of QCA are briefly introduced and invented data are used to illustrate how it may be used to integrate cross-case and within-case analysis in research settings.

This chapter is deliberately placed at the end of the book, as it is a fitting conclusion to the preceding chapters in Part 5 on quantitative and qualitative data analysis, showing how researchers can move beyond such dichotomous thinking. As its title suggests, the discussion moves beyond mixed methods approaches to examine QCA for integrating cross-case and within-case analyses, addressing:

45.1 Introduction

Mixing methods refers to combining quantitative and qualitative research in one study (see Chapter 2). As Small (2011) and Harding and Seefeldt (2013) note, the literature on mixed methods is enormous and growing, though mixing methods is hardly new (see, for example, Lacey, 1970). While we do not accept that the qualitative/quantitative divide is as clear-cut as some believe (Cooper *et al*., 2012), we should briefly note what these terms usually denote. Quantitative research *typically* focuses on cross-case analysis of large data sets, aiming to produce via correlation-based techniques generalizable knowledge concerning the relative importance of independent variables in explaining some dependent variable. The details of individual cases, or even types of case, tend to get lost during statistical aggregation (though see cluster analysis (Cooper and Glaesser, 2011) and latent class analysis (Vermunt and Magidson, 2002) for examples of quantitative techniques that do more to preserve the holistic case). Qualitative work takes many forms. It can also involve cross-case analysis, though usually of a smaller number of non-randomly selected cases. It is more likely than quantitative work to involve within-case analysis, focusing on interpretative, processual and narrative analysis. Rather than on supposedly independent variables, it tends to focus on the holistic case as the active agent.

In his review of mixed methods studies, Small (2011) employs an overarching distinction between studies mixing methods of *data collection* and those mixing forms of *data analysis*. He notes that mixed forms of data might be used primarily either to confirm findings or to complement each other's strengths and weaknesses, that these data might be collected sequentially or in parallel, and in either a nested or non-nested manner (where nesting involves choosing cases for in-depth analysis from the sample used in the large-n study). Regarding analyses, he discusses a range of types of combination. Discussing one, *integrative analyses*, he notes that a small number of researchers have integrated analyses by creating new techniques, adding, 'without a doubt, the most successful of these has been Ragin's (1987, 2000, 2008) qualitative comparative analysis (QCA)' (p. 77).

Small also discusses how mixing methods can help establish causal claims (see also Chapter 6). Harding and Seefeldt (2013) focus specifically on causal analysis. Assuming a backdrop of large-scale quantitative studies, they note the multiple roles which a qualitative component can play in causal inference, and discuss research design decisions arising when quantitative and qualitative methods are to be integrated. They

discuss the important distinction between cross-case and within-case analysis within qualitative research seeking causal knowledge, commenting that Ragin's set theoretic QCA is well suited to the former. However, having noted that it is conventional (correlation-based) quantitative methods that are more usually employed to undertake cross-case analysis in mixed methods studies, they choose to concentrate on discussing the logic of *within*-case analysis in qualitative research, arguing, with Mahoney (2000), that within-case analysis in the form of *process-tracing* and *pattern matching* can make important contributions to causal analysis. They provide a strong case that qualitative work can offer much to quantitative researchers seeking causal knowledge. They conclude that empirical studies should therefore more often include a mixed methods approach in their design.

We share this view, but in this chapter we will focus specifically on what Small (2011) claims is the most successful new integrative technique: Ragin's (1987, 2000, 2008) QCA. Ragin (1994b) argues that QCA, a Boolean analytic approach drawing on logic and set theory, can bridge the wide 'methodological gulf between intensive case-oriented research and extensive variable-oriented research' (p. 304), suggesting that it will be an appropriate approach for those willing to mix methods. This bridging requires 'tools that preserve the intensity of the case-oriented approach, especially its attention to combinations and configurations of causes and conditions, when examining many cases' (p. 304).

Ragin (2004, 2006a) notes that conventional linear quantitative methods typically aim to assess the relative importance of independent variables in predicting some outcome, i.e. to assess the net effects of variables having controlled for others. A well-known example is the debate between Breen and Goldthorpe (1999, 2002) and Saunders (1997) concerning the relative importance of social class and ability in explaining educational achievement (see also Chapter 6). Such work is clearly important. However, there is an alternative, and perhaps more fruitful, way of conceptualizing such questions. This, employing a model of conjunctural causation, draws on the concepts of necessary and sufficient conditions. We might hypothesize, for example, that, for individuals from higher class origins, high ability will tend to be sufficient but not necessary for later high educational achievement while, for lower class respondents, high ability will tend to be necessary but not sufficient (Glaesser and Cooper, 2011; Cooper *et al*., 2012). Such an approach can be developed in more complex ways. It might be, for example, that the conjunction 'being male, from a higher class origin, and of a certain level of ability, but *not* from ethnic background X' is sufficient for some outcome. QCA facilitates such cross-case analyses. QCA allows the cross-case component of a mixed methods study to focus more on the holistic *case* than do conventional quantitative techniques such as regression analysis. Not only is this, for us, a good thing in itself, but it also allows a theoretically coherent form of sequential integration of cross-case and within-case analyses.

In the next section we consider why researchers employing a broadly quantitative approach, but wishing to develop causal knowledge, might need to introduce a qualitative component into their research design. In the subsequent section, we reverse the direction of argument, and consider why a qualitative researcher wishing to undertake a small number of comparative case studies would be well-advised to introduce some rigorous mathematical considerations – of a particular type – into his/her research planning. We then illustrate the use of Ragin's QCA as one way in which quantitative and qualitative approaches may be integrated. We conclude by considering the implications of our discussion.

45.2 Starting from a 'quantitative' stance

We assume here that researchers prefer causal knowledge. This is not because descriptive knowledge is not useful. The fact, for example, that educational achievement varies by social class is important to know, but we ideally want the link explained. The boundary between description and explanation is not, however, clear-cut (see also Chapter 4). While Merton (1987) notes that ' "establishing the phenomenon" involves the doctrine … that phenomena should of course be shown to exist or to occur before one explains why they exist or how they come to be' (p. 1), this is not always straightforward. In the case of our earlier example, one needs to consider possible explanations for the class–achievement link before accepting it as a useful description. Perhaps cognitive capacity and/or attitudinal differences underlie the link and, once these are controlled, the class–achievement link is weakened. These considerations already move us beyond description and towards explanation.

We can also note that 'social class' is a summarizing variable, with social, economic, relational and cultural elements. Even having decided that the class–achievement link is not an artefact of our having omitted important factors from our analysis, we will still want to know what aspect of class produces the link. Or consider the more applied field of educational

evaluation. If a new pedagogic technique, in some randomized controlled trial (RCT), doesn't produce the average magnitude of gains that its developers had hoped for, we would want to understand why (see also Chapter 20). We would want, for example, to know something about the implementation process. Perhaps the technique was not actually introduced in classrooms because most teachers believed it to be a distraction from preparing students for tests. Or perhaps it required skills that teachers lacked. Without researching implementation it would be premature to report that the technique had no potential value. It would be especially useful to gain comparative knowledge about the classrooms in which it had 'worked'.

For these reasons, we believe that research in the social and educational field is usually more valuable when it delivers more than description, whether in words or equations, of the pattern of regularities that exist in some data set (such as 'higher' class → 'higher' achievement). There are, of course, scholars who argue that cross-case work – the analysis of regularities – can generate causal knowledge, at least in ideal circumstances. We would agree that an RCT, in ideal circumstances, can provide good grounds for accepting that some intervention has caused an average effect in some particular time and place. Similarly, in ideal circumstances, the quantitative analysis of regularities, via correlational or other techniques, can provide knowledge of the relative importance of the factors causing some outcome (Spirtes *et al*., 2001; Baumgartner, 2008; Pearl, 2009). In practice, there are many threats to the validity of such work (Lieberson, 1985; Freedman, 1991; Morrison, 2001). Here we will just briefly note two arguments that aim to qualify the claims made by the advocates of RCTs and survey techniques, and that provide additional grounds for combining qualitative work with quantitative analysis.

Discussing RCTs, Cartwright and Hardie (2012) agree that an RCT can show that some intervention worked in some particular time and place. Policy makers, however, want to know whether the same intervention will work in new settings. The authors stress that an RCT actually shows that an intervention worked *in conjunction with many other support factors*. Unless these, or substitutes, are present in the new settings we have no grounds for expecting the results of the RCT to be transferable. An example often used concerns class size. Even if an RCT shows that halving class sizes produces learning gains, we would need to know before halving all the classes in our country that there were enough *relevantly skilled* teachers available. While there may have been enough of these to staff the additional trial classes, we may lack this relevant support factor when scaling up. To gain knowledge about support factors, we would need to draw both on established theoretical understanding and to undertake some in-depth study of the processes of implementation of the intervention in context. Here is another compelling reason for combining case studies with quantitative analysis.

Analogous problems to those noted by Cartwright and Hardie apply to attempts to gain causal knowledge in social and educational settings via quantitative survey studies employing correlational techniques. Lieberson (1985) provides a thorough analysis of many of these. The coefficients in a regression equation provide an algebraic snapshot of the relationships between a set of independent variables and a dependent variable at some time and place, given the particular choice of variables in the model (see Chapter 42). They will also reflect existing relations between competing interest groups and between these and existing societal laws and regulations.

A regression equation may tell us that, having controlled other factors, a large part of the difference in earned income between ethnic groups A and B (with group A earning a higher income) is statistically 'explained' by prior differences in educational achievement. Policy makers may then decide that the route to improved income for members of group B is improved achievement. Measures are introduced to achieve this goal. However, members of group A, unsurprisingly, wish to defend their position. They use their existing resources to enhance the educational careers of their children by employing private tutors. They also use their social and occupational contacts to arrange internships for their children. As a result, the expected income gain for group B does not occur. Basic causes here, as is often the case, override more superficial ones (Lieberson, 1985). Once again, we think, in order to understand what are more basic and what are more superficial causes requires more than just the analysis of regularities that is usually delivered by quantitative research methods. Process-tracing (George and Bennett, 2005) through case studies, in conjunction with the use of established theoretical knowledge, offers, as we noted earlier, a way forward.

So far, we have considered what a qualitative component might add to quantitative studies. Now we reverse the line of argument. Why might a qualitative researcher benefit from some of the mathematical thinking that characterizes quantitative work? What type of mathematics might be most useful?

45.3 Starting from a 'qualitative' stance

In the previous section, we discussed situations in which a quantitative study might usefully be supplemented by qualitative research. It may also make sense for qualitative work to precede a quantitative study, for example when qualitative work is used in an exploratory, hypothesis-generating way and any initial findings are then confirmed and elaborated on by conducting a larger-scale study (e.g. Cooper, 1998; Cooper and Dunne, 2000). There are also other situations in which causally orientated qualitative work may benefit from the addition of a quantitative element, and we discuss one of these next.

Consider a postgraduate research student wanting to undertake case studies of individuals' educational careers, situating these within personal and social contexts in order to explain why some individuals, but not others, attend university. She has read widely and, on the basis of this reading, proposes a small number of key explanatory factors. These are social class, ethnicity, gender and early 'academic ability'. She takes a configurational approach to causation, claiming that the effect of any factor will depend on other characteristics of the individual (see also Chapter 6). For example, in order to reach university, high 'ability' may be necessary for individuals from some class and ethnic backgrounds but moderate 'ability' might be sufficient for individuals from other backgrounds, given the use of private tutors.

She also thinks that, given the intensity of case studies, she might manage twenty case studies. She intends to select cases to cover the various combinations of characteristics, exploring the processes by which they have their effects. In order to keep things manageable, she proposes using a simplified threefold class scheme and using a simple division of 'ability' into high, moderate and low categories. The society under study has a majority ethnic group (A) and two minority groups (B, C). A simple binary gender division is also proposed. The student has not calculated the number of possible distinct types of case this categorization generates. It is $3 \times 3 \times 3 \times 2$, i.e. 54. The supervisor explains this, pointing out that, even were just one case per type to be explored, the range of types cannot be covered with twenty cases.

The student wants, however, to study the whole range of types in order to explore the interaction of the full ranges of the various characteristics in producing the outcome, and also believes that only via case studies can she gain access to the chains of processes that produce the correlations that she has seen in the statistical literature. The supervisor sees such complete coverage as unrealistic and suggests that she will need to think about the selection of cases for in-depth study. She should also consider whether her study should include a quantitative element – in the sense of a larger n – to allow coverage of the types of cases that her case studies will not cover.

The research supervisor ascertains that the student has been reading regression-based quantitative studies and narrative-based qualitative accounts of educational careers (see Chapters 42 and 35 respectively). She has come across discussions of how to integrate regression-based studies with case studies (Lieberman, 2005; Rohlfing and Starke, 2013) and also much debate about mixed methods, but has not become aware of the tradition established by researchers who, while favouring case-based approaches, also concern themselves with what she has come to fear are 'positivist' concerns such as sampling, proof and rigorous logical analysis (Becker, 1958) (see Chapter 1). Referring to Becker's discussion of logic in his *Tricks of the Trade* (1998) and Ragin's QCA, the supervisor suggests to the student that she take seriously the idea of combining Boolean cross-case analyses of regularities and within-case analysis of processes in order to explore in a *configurational* manner her chosen research theme. This will allow her to develop, he claims, good grounds for choosing cases to develop her understanding of the complex processes she wishes to understand. With this plan in mind, we now introduce QCA.

45.4 Qualitative Comparative Analysis (QCA)

QCA can be used with 'crisp sets' representing binary conditions such as male/female, with multivalued conditions (Cronqvist and Berg-Schlosser, 2009) such as high/moderate/low social class, or by employing 'fuzzy sets' with fully continuous conditions such as measured ability (Ragin 2000, 2008). Since our focus is the core features of integrating cross-case and within-case analysis via QCA, we restrict ourselves to binary conditions. We should note that there are various free software packages available to undertake QCA analyses (e.g. Ragin and Davey, 2014; Duşa, 2016; Thiem, 2016).

QCA employs set theory as the basis for the Boolean analysis of conjunctural causation (Ragin, 1987, 2000, 2006b, 2008). It is important to note that this fundamentally differentiates QCA and related techniques such as coincidence analysis (CNA) (Baumgartner, 2009) from conventional correlation-based analysis

(Thiem *et al*., 2016). In QCA, cases are seen as members of various sets defined by conditions such as 'being male', 'having high ability', 'achieving a degree'. For a condition to be strictly sufficient, logically, for an outcome, all members of the condition set must also be members of the outcome set. This subset relation is equivalent to the condition logically implying the outcome, as it does in Table 45.1, where all cases with 'high ability' achieve a degree. Note though that 'high ability' is not necessary for the outcome here, since 5 per cent of those lacking 'high ability' achieve a degree (perhaps by purchase in a corrupt state). For a condition to be strictly necessary, logically, for an outcome, we need the outcome set to be a subset of the condition set, i.e. all cases with the outcome must be members of the conditions set, as is the case in Table 45.2. Here every case with a degree also has 'high ability' but it is not the case that 'high ability' is sufficient for achieving a degree.

These are simple examples. We must note four things. *First*, the condition can be more complex, for example, a conjunction of factors such as MALE*HIGH ABILITY*HIGH CLASS where * denotes set intersection (logical AND). Here a case must be a member of all three sets to belong to the *configuration*.

Second, given that the empirical social world is less tidy than Tables 45.1 and 45.2, we might relax the criterion for sufficiency, either by referring to quasi-sufficiency where, say, 80 per cent of those with a simple or complex condition achieve the outcome (Ragin, 2000), or by using the proportion of those with a condition who achieve the outcome as a measure of the *consistency* of the subset relation with strict sufficiency (Ragin, 2006b). Where a condition is considered sufficient, it is also possible to assess how many of the cases with the outcome it explains, via the concept of coverage (Ragin, 2006b). Similar procedures can be applied to necessity.

TABLE 45.1 DATASET WHERE CONDITION IS SUFFICIENT BUT NOT NECESSARY FOR THE OUTCOME

Condition 'high ability'	*Outcome 'achieves degree'*	
	Absent	*Present*
Present	0	2000
Absent	1900	100

TABLE 45.2 DATASET WHERE CONDITION IS NECESSARY BUT NOT SUFFICIENT FOR THE OUTCOME

Condition 'high ability'	*Outcome 'achieves degree'*	
	Absent	*Present*
Present	1000	1000
Absent	2000	0

Third, QCA uses Boolean minimization to simplify its analyses (Ragin, 1987). For example, were the conjunctions A*B*C and A*B*c (where upper case letters indicate the presence of a condition and lower case its absence) both found to be sufficient for the outcome Y, then, since the presence or absence of C makes no difference to whether the outcome is achieved, QCA will drop it, and say that the conjunction A*B is sufficient (though see Cooper and Glaesser (2012), for some discussion of what this minimization procedure might hide).

Last, it is sometimes the case, especially when the data set is small, that there may be no cases for some conjunctions. Such limited diversity raises difficulties for Boolean analysis since, to use the previous example, were we to have cases of A*B*C all achieving the outcome Y, but no cases of A*B*c in our dataset, we would have to reflect carefully – and counterfactually – about whether cases lacking C, were they to exist, would achieve Y. If we thought they would, then we could reduce A*B*C to A*B. If not, we would have to retain A*B*C as our solution. It is important to note that there is considerable debate concerning such counterfactual reasoning. Baumgartner (2009) has developed an alternative form of minimization as part of his Coincidence Analysis (CNA) that does not, when certain assumptions are met, require pairs like A*B*C and A*B*c. In addition, the use of counterfactual reasoning to allow minimisation beyond what the empirical data set would seem to warrant is also contentious (see Schneider and Wagemann 2012, 2015; Thiem, 2015; Cooper and Glaesser, 2016a).

We next discuss an invented example to illustrate how integration of techniques using QCA can be undertaken. This will also allow us to introduce 'truth tables', a key feature of QCA. To keep our example simple, we employ a two-fold class scheme (higher class of origin=1, lower=0), gender (male=1, female=0), two ethnic groups (majority group=1, minority=0) and measured academic ability (high=1, not high=0). We assume that there is a dataset available (n=2,160) with information on these and other factors.

45.5 QCA: sufficiency

Recall our research student who is focusing on what is sufficient for achieving admission to university and who believes that class and ability are two of the key factors. She begins by undertaking a simple Boolean cross-case analysis of the relation between entering university (U) and high-class origins (HC) and high ability (HA). Table 45.3, termed a 'truth table' because of its similarities with logicians' truth tables, provides a representation of the cross-case relations in a data set that, while containing just the same information as a complex crosstabulation, enables us to see more easily the 'types' of cases that are our concern and which of these types achieve the outcome. In the context of QCA, we think of the rows of the table as sets of cases, but also as types. Row 1, for example, comprises cases of high ability from high-class origins. We assume, for simplicity, that our sample accurately represents some population (see Cooper and Glaesser (2016b) and Thiem *et al*. (2016) on alternative scenarios). We use an illustrative threshold of 80 per cent of cases achieving the outcome to test for quasi-sufficiency. On this basis, we allocate a 1 in the U (outcome) column for any rows where the consistency proportion is at least 0.8, and a zero for other rows. It can be easily seen that the configuration HA*HC is (logically) quasi-sufficient for the outcome, since 95 per cent of these cases enter university (though only approximately 34.5 per cent of cases with the outcome U are 'covered' by HA*HC).

Our researcher is not satisfied, however, with this Boolean cross-case equivalent of a mere correlation and wants to develop an understanding of the mechanisms and processes that explain (causally) the link between HA*HC and university. For this purpose, she plans to make use of in-depth interviews with a small number of cases in order to explore, via process-tracing, what it is about class and ability that explain the patterns in Table 45.3. Considering class, for example, is it cultural or economic features of class, or both, that explain the link? Or can class patterns be explained by rational choice theory (Breen and Goldthorpe, 1997) (see Chapter 6)? To explore these questions via within-case exploration, she selects cases from row 1 that achieve the outcome. Our researcher also wants to understand what it is about the 5 per cent of cases in row 1 who do not achieve the outcome that explains this fact. She therefore selects from row 1 some cases for interview who have not achieved the outcome.

We shall also assume that this researcher would like, as part of theory development and testing, to improve the 0.95 consistency figure. She wants to find a configuration of factors that is nearer perfect sufficiency. In the context of sufficiency, as Ragin and Schneider (2011) note, theory development requires us to try to raise consistency measures by adding factors to the configurations that form the rows of truth tables. The basic idea is to move from X1 to 'X1 combined with X2'. If we add a single dichotomous factor X2 then each row of the truth table will be split into two, doubling the size of the table, and any causal arguments will be more detailed, i.e. less inclusive (p. 159). The question is, how should we find candidate X2s to add to our QCA cross-case analysis? Established theory will help, but so will in-depth case study. Our researcher, as a result of the interviewing of selected cases, alongside establishing what explains the *typical* cases in row 1 (who achieve the outcome), begins to suspect that the other two factors she was initially interested in, sex and ethnicity, indeed interact with class and ability in determining whether the outcome is achieved. In particular, she suspects that the combination of being an ethnic minority female with HA and HC will cover most of the *deviant* cases in row 1 – those who, atypically, do not achieve the outcome.

With this in mind, she returns to the data set to undertake another round of cross-class analysis, but this time incorporating the additional factors of sex and ethnicity. This produces the truth table in Table 45.4. Row 1 of Table 45.3 is now split into rows 1–4 of this table.

TABLE 45.3 TRUTH TABLE FOR U = F(HA, HC), USING 0.8 THRESHOLD FOR CONSISTENCY WITH SUFFICIENCY

Row	*HA*	*HC*	*number*	*Number entering university*	*U*	*Consistency*
1	1	1	220	209	1	0.95
2	0	1	240	110	0	0.458
3	1	0	600	255	0	0.425
4	0	0	1100	32	0	0.029

TABLE 45.4 FULL TRUTH TABLE FOR U = F(HA,HC,ME,M), USING 0.8 THRESHOLD FOR CONSISTENCY WITH SUFFICIENCY

Row	*High ability*	*High origin class*	*Majority ethnic*	*Male*	*Number of cases*	*Achieves university*	*Does not achieve university*	*Consistency for sufficiency re outcome*	*Quasi-sufficient for outcome*	*Quasi-sufficient for not outcome*	*Quasi-sufficient for neither*
1	1	1	1	1	100	99	1	0.990	1	0	0
2	1	1	1	0	100	99	1	0.990	1	0	0
3	1	1	0	1	10	9	1	0.900	1	0	0
4	1	1	0	0	10	2	8	0.200	0	1	0
5	1	0	1	1	200	100	100	0.500	0	0	1
6	1	0	1	0	200	100	100	0.500	0	0	1
7	1	0	0	1	100	50	50	0.500	0	0	1
8	1	0	0	0	100	5	95	0.050	0	1	0
9	0	1	1	1	100	50	50	0.500	0	0	1
10	0	1	1	0	100	50	50	0.500	0	0	1
11	0	1	0	1	20	10	10	0.500	0	0	1
12	0	1	0	0	20	0	20	0.000	0	1	0
13	0	0	1	1	500	20	480	0.040	0	1	0
14	0	0	1	0	500	10	490	0.020	0	1	0
15	0	0	0	1	50	2	48	0.040	0	1	0
16	0	0	0	0	50	0	50	0.000	0	1	0

Rows 1 and 2 of Table 45.4, where HA and HC are combined with being a member of the majority ethnic group, now have improved consistencies with sufficiency of 0.99. Whether the case is male or female makes no difference in this context. However, the consistencies of rows 3 and 4 have dropped below 0.95. In row 4, where HA and HC are combined with being a female from the minority ethnic group, only 20 per cent achieve the outcome.

The researcher, wanting to offer policy advice on how to improve the situation of this group, decides to carry out further in-depth case studies of cases with HA*HC*me*m. What is it about being a female from a minority ethnic background that explains this low percentage, given the positive class and ability factors? Similar questions can be asked about the other rows of our initial Table 45.3. What is it that explains the small proportion from row 4 who do manage, against the odds, to achieve the outcome? In rows 2 and 3, where only one of HA or HC is present, what determines whether a case does or does not achieve the outcome?

Similar moves between cross-case and within-case analysis can be undertaken in relation to the analysis of necessary conditions (Glaesser and Cooper, 2011; Schneider and Rohlfing, 2013; Glaesser, 2015). The disjunction 'HA or HC' is quasi-necessary for the achievement of Y, for example, with a consistency with necessity of approximately 0.95. There are, nevertheless, thirty-two cases who achieve U without being either HA or HC. Our researcher might decide to interview such cases with the outcome (from row 4 of Table 45.3, or rows 13–16 of Table 45.4), in order to explore how this occurs. She may find that type of ethnic minority background matters, and then return to her original categorization of two minority groups in addition to the ethnic majority.

45.6 Conclusion

In this chapter, following a brief discussion of the nature of mixed methods research (see also Chapter 2), we have presented one way of integrating within-case and cross-case analyses. We have argued for the use of QCA as a means of integration not only because we have found this useful in our own work but because we believe its case-based Boolean approach allows it to offer rigour to researchers who wish to explore configurational causation.

We end this chapter by pointing to some other relevant literature. For researchers who wish to explore the combining of case studies with more conventional techniques, there are many useful sources of advice (e.g. Seawright and Gerring, 2008). We have chosen, given

constraints of space, to use an invented example. For our application of the procedures we have described to real data, see Cooper and Glaesser (2012) and Glaesser (2015) where in-depth interviews are employed as the qualitative component of a study of educational transitions in England and Germany. For another example that combines the use of case study with Boolean cross-case analysis, see Berg-Schlosser (2012). For discussion of case selection when fuzzy set QCA is employed, see Schneider and Rohlfing (2016).

Companion Website

The companion website to the book provides PowerPoint slides for this chapter, which list the structure of the chapter and then provide a summary of the key points in each of its sections. This resource can be found online at: **www.routledge.com/cw/cohen**.

Bibliography

Aaen, J. and Dalsgaard, C. (2016) Student Facebook groups as a third space: between social life and schoolwork. *Learning, Media and Technology*, 41 (1), pp. 160–86.

Abascal, E. and Diaz de Rada, V. (2014) Analysis of 0 to 10-point response scales using factorial methods: a new perspective. *International Journal of Social Research Methodology*, 17 (5), pp. 569–84.

Adair, J. K. and Pastori, G. (2011) Developing qualitative coding frameworks for educational research: immigration, education and the Children Crossing Borders project. *International Journal of Research and Method in Education*, 34 (1), pp. 31–47.

Adelman, C., Kemmis, S. and Jenkins, D. (1980) Rethinking case study: notes from the Second Cambridge Conference. In H. Simons (ed.) *Towards a Science of the Singular*. Norwich, UK: Centre for Applied Research in Education, University of East Anglia, pp. 45–61.

Adler, P. A. and Adler, P. (1994) Observational techniques. In N. K. Denzin and Y. S. Lincoln (eds) *Handbook of Qualitative Research*. London: Sage, pp. 377–92.

Agar, M. (1993) Speaking of ethnography. Cited in D. Silverman (1993) *Interpreting Qualitative Data*. London: Sage, pp. 520–30.

Agee, J. (2009) Developing research questions: a reflective process. *International Journal of Qualitative Studies in Education*, 22 (4), pp. 431–47.

Agell, L., Soria, V. and Carrió, M. (2015) Using role-play to debate animal testing. *Journal of Biological Education*, 49 (3), pp. 309–21.

Aiken, L. R. (2003) *Psychological Testing and Assessment* (eleventh edition). Boston, MA: Pearson Education Group Inc.

Airasian, P. W. (2001) *Classroom Assessment: Concepts and Applications* (fourth edition). New York: McGraw-Hill.

Akbulut, Y. (2015) Predictors of inconsistent responding in web surveys. *Internet Research*, 25 (1), pp. 131–47.

Alderson, P. and Morrow, V. (2011) *The Ethics of Research with Children and Young People: A Practical Handbook*. London: Sage.

Aldridge, A. and Levine, K. (2001) *Surveying the Social World: Principles and Practice in Survey Research*. Buckingham, UK: Open University Press.

Aldridge, J. M. and Fraser, B. J. (2000) A cross-cultural study of classroom learning environments in Australia and Taiwan. *Learning Environments Research*, 3 (2), pp. 101–34.

Aldridge, J. M., Fraser, B. J. and Huang, T.-C. I. (1999) Investigating classroom environments in Taiwan and Australia with multiple research methods. *The Journal of Educational Research*, 93 (1), pp. 48–62.

Alexander, R. J. (2000) *Culture and Pedagogy: International Comparisons in Primary Education*. Oxford: Basil Blackwell.

Allison, P. D. (2001) *Missing Data*. Thousand Oaks, CA: Sage.

Altricher, H. and Gstettner, P. (1993) Action research: a closed chapter in the history of German social science? *Educational Action Research*, 1 (3), pp. 329–60.

Alvesson, M. and Sandberg, J. (2011) Generating research questions through problematization. *Academy of Management Review*, 36 (2), pp. 247–71.

Alvesson, M. and Sandberg, J. (2013) *Constructing Research Questions: Doing Interesting Research*. London: Sage.

American Educational Research Association (2000) *Ethical Standards of the American Educational Research Association 2000*. Washington, DC: American Educational Research Association. Available from: www.aera.net/uploadedFiles/About_AERA/Ethical_Standards/EthicalStandards.pdf [Accessed 17 April 2010].

American Educational Research Association (2011) Code of ethics. *Educational Researcher*, 40 (3), pp. 145–56.

American Psychological Association (2002) *Ethical Principles and Code of Conduct*. Available from: www.apa.org/ethics/code2002.html [Accessed 15 May 2005].

American Psychological Association (2010) *Publication Manual of the American Psychological Association*. Washington, DC: American Psychological Association.

American Psychological Association (2014) *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

American Sociological Association (1999) *Code of Ethics and Policies and Procedures of the ASA Committee on Professional Ethics*. Available from: www.asanet.org/members/ecoderev.html [Accessed 15 May 2005].

Anderson, B. (1983) *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London: Verso.

Anderson, D. S. and Biddle, B. J. (eds) (1991) *Knowledge for Policy: Improving Education through Research*. London: Falmer.

Anderson, G. and Arsenault, N. (1998) *Fundamentals of Educational Research* (second edition). London: Routledge Falmer.

Anderson, L. (2006) Analytic autoethnography. *Journal of Contemporary Ethnography*, 35 (4), pp. 373–95.

Anderson, N., Schlueter, J. F., Carlson, J. F. and Geisinder, J. F. (2016) *Tests in Print IX*. Lincoln, NE: University of Nebraska Press.

Anderson, T. and Shattuck, J. (2012) Design-based research: a decade of progress in education research? *Educational Researcher*, 41 (1), pp. 16–25.

Andrelchik, H. (2016) Success is cheesecake: a guide to analysing student discourse. *International Journal of Qualitative Studies in Education*, 29 (2), pp. 135–49.

Andrews, R. (2003) *Research Questions*. London: Continuum.

Anfara, V. A., Brown, K. M. and Mangione, T. L. (2002) Qualitative analysis on stage: making the research process more public. *Educational Researcher*, 31 (7), pp. 28–38. Available from: http://35.8.171.42/aera/pubs/er/pdf/vol. 31_07/AERA310706.pdf [Accessed 29 October 2005].

Angrist, J. D. (2003) Randomized trials in quasi-experiments in education research. *NBER Reporter Research Summary*. Cambridge, MA: National Bureau of Economic Research.

Arabacioglu, T. and Ajar-Vural, R. (2014) Using Facebook as a LMS? *The Turkish Online Journal of Educational Technology*, 13 (2), pp. 202–14.

Archer, T. M. (2003) Web-based surveys. *Journal of Extension*, 41 (4), article 4TOT6. Available from: www.joe.org/joe/2003august/tt6.php [Accessed 24 March 2016].

Arditti, J. A. (2002) Doing family research at the jail: reflections of a prison widow. *The Qualitative Report*, 7 (4). Available from: www.nova.edu/ssss/QR/QR7-4/arditti.html [Accessed 21 November 2003].

Argyle, M. (1978) Discussion chapter: an appraisal of the new approach to the study of social behaviour. In M. Brenner, P. Marsh and M. Brenner (eds) *The Social Contexts of Method*. London: Croom Helm, pp. 237–55.

Argyris, C. (1958) Review of 'Supervisory and Executive Development. A Manual for Role Playing'. *Management Science*, 4 (3), pp. 321–2.

Argyris, C. (1990) *Overcoming Organizational Defenses: Facilitating Organizational Learning*. Boston: Allyn & Bacon.

Arksey, H. and Knight, P. (1999) *Interviewing for Social Scientists*. London: Sage.

Arnold, R. (1998) The drama in research and articulating dynamics. In J. Saxton and C. Miller (eds) *The Research of Practice: The Practice of Research*. Victoria, BC: International Drama in Education Research Institute, pp. 110–31.

Arnon, S. and Reichel, N. (2009) Closed and open questions tools in a telephone survey about 'the good teacher'. *Journal of Mixed Methods Research*, 3 (2), pp. 172–96.

Aronowitz, S. and Giroux, H. A. (1991) *Postmodern Education: Politics, Culture, and Social Criticism*. Minneapolis, MN: University of Minnesota Press.

Aronson, E. and Carlsmith, J. M. (1969) Experimentation in social psychology. In G. Lindzey and E. Aronson (eds) *The Handbook of Social Psychology, Vol. 2* (second edition). Reading, MA: Addison-Wesley, pp. 1–79.

Arsenault, N. and Anderson, G. (1998) Qualitative research. In G. Anderson and N. Arsenault (eds) *Fundamentals of Educational Research* (second edition). London: Routledge Falmer, pp. 119–35.

Arthur, J., Waring, M., Coe, R. and Hedges, L. V. (eds) (2012) *Research Methods and Methodologies in Education*. London: Sage.

Arthur, L. and Cox, E. (2014) From evaluation to research. *International Journal of Research and Method in Education*, 37 (2), pp. 137–50.

Ary, D., Jacobs, L. C. and Razavieh, A. (2002) *Introduction to Research in Education* (sixth edition). Belmot, CA: Wadsworth/Thomson Learning.

Ary, D., Jacobs, L. C., Razavieh, A. and Sorensen, C. (2006) *Introduction to Research in Education* (seventh edition). Belmont, CA: Wadsworth.

Ashraf, H., Motlagh, F. G. and Salami, M. (2014) The impact of online games on learning English vocabulary by Iranian (low-intermediate) EFL learners. *Procedia – Social and Behavioral Sciences*, 98, pp. 286–91.

Association of Internet Researchers (2012) *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0)*. Available from: http://aoir.org/reports/ethics2.pdf [Accessed 2 February 2016].

Atkinson, J. M. and Heritage, J. (1999) Transcript notation: structures of social action – studies in conversation analysis. *Aphasiology*, 13 (4), pp. 243–9.

Atkinson, P. (1997) Narrative turn or blind alley? *Qualitative Health Research*, 7 (3), pp. 325–44.

Atkinson, P. (2006) Rescuing autoethnography. *Journal of Contemporary Ethnography*, 35 (4), pp. 400–4.

Atkinson, P. and Delamont, S. (2006) In the roiling smoke: qualitative inquiry and contested fields. *International Journal of Qualitative Studies in Education*, 19 (6), pp. 747–55.

Atkinson, R. (1998) *The Life Interview*. London: Sage.

Auerbach, C. F. and Silverstein, L. B. (2003) *Qualitative Data: An Introduction to Coding and Analysis*. New York: New York University Press.

Austin, J. L. (1962) *How to Do Things with Words*. Oxford: Oxford University Press.

Axline, V. (1964) *Dibs In Search of Self*. New York: Ballantine.

Babbie, E. R. (2010) *The Practice of Social Research* (eleventh edition). New York: Thompson.

Bacharach, S. B. (1989) Organizational theories: some criteria for evaluation. *Academy of Management Review*, 14 (4), pp. 496–515.

Bailey, K. D. (1994) *Methods of Social Research* (fourth edition). New York: The Free Press.

Bailey, K. D. (2007) *Methods of Social Research* (fifth edition). New York: The Free Press.

Bair, C. R. (1999) Meta-synthesis. Paper presented at the 24th annual meeting of the Association for the Study of Higher Education, San Antonio, 18–21 November.

Bak, P. (1996) *How Nature Works*. New York: Copernicus.

Bakardjieva, M. and Feenberg, A. (2000) Involving the virtual subject. *Ethics and Information Technology*, 2 (4), pp. 233–40.

Baker, B. (1999) What is voice? Issues of identity and representation in the framing of reviews. *Review of Educational Research*, 69 (4), pp. 365–83.

Baker, T. L. (1994) *Doing Social Research* (second edition). New York: McGraw-Hill.

Bakhtin, M. (1981) Discourse in the novel. In M. Holquist (ed.) *The Dialogic Imagination*. Austin, TX: University of Texas Press, pp. 259–422.

Baldwin, R. G. and Austin, A. E. (1995) Toward greater understanding of faculty research collaboration. *Review of Higher Education*, 19 (1), pp. 45–70.

Ball, S. J. (1990) *Politics and Policy Making in Education*. London: Routledge.

Ball, S. J. (1994a) *Education Reform: A Critical and Post-Structuralist Approach*. Buckingham, UK: Open University Press.

Ball, S. J. (1994b) Political interviews and the politics of interviewing. In G. Walford (ed.) *Researching the Powerful in Education*. London: UCL Press, pp. 96–115.

Bampton, R. and Cowton, C. J. (2002) The e-interview. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 3 (2), pp. 1–12, article 4. Available from: http://nbnresolving.de/urn:nbn:de:0114-fqs020295 [Accessed 28 March 2010].

Banham, M. (ed.) (1995) *The Cambridge Guide to Theatre* (second edition). Cambridge: Cambridge University Press.

Banks, M. (1995) Visual research methods. *Social Research Update*, 11, pp. 1–6. Available from: http://sru.soc.surrey.ac.uk/sru11/sru11.html [Accessed 10 May 2010].

Banks, M. (2007) *Using Visual Data in Qualitative Research*. London: Sage.

Bannister, D. (ed.) (1970) *Perspectives in Personal Construct Theory*. London: Academic Press.

Bannister, D. and Mair, J. M. M. (1968) *The Evaluation of Personal Constructs*. London: Academic Press.

Banuazizi, A. and Movahedi, A. (1975) Interpersonal dynamics in a simulated prison: a methodological analysis. *American Psychologist*, 30 (2), pp. 152–60.

Banville, D., Desrosiers, P. and Genet-Volet, Y. (2000) Translating questionnaires and inventories using a cross-cultural translation technique. *Journal of Teaching in Physical Education*, 19 (3), pp. 374–97.

Bargh, J. A., McKenna, K. Y. A. and Fitzsimons, G. M. (2002) Can you see the real me? Activation and expression of the 'true self' on the Internet. *Journal of Social Issues*, 58 (1), pp. 33–48.

Barker, C. D. and Johnson, G. (1998) Interview talk as professional practice. *Language and Education*, 12 (4), pp. 229–42.

Barley, R. and Bath, C. (2014) The importance of familiarisation when doing research with young children. *Ethnography and Education*, 9 (2), pp. 182–95.

Barnes, N., Penn-Edwards, S. and Sim, C. (2015) A dialogic about using Facebook status updates for education research: a PhD student's journey. *Educational Research and Evaluation*, 21 (2), pp. 109–21.

Barnes, S. B. (2004) Issues of attribution and identification in online social research. In M. D. Johns, S. S. Chen and G. J. Hall (eds) *Online Social Research: Methods, Issues and Ethics*. New York: Peter Lang, pp. 203–22.

Barnett-Page, E. and Thomas, J. (2009) Methods for the synthesis of qualitative research: a critical review. *BMC Medical Research Methodology*, 9 (1), p. 59.

Baron, R. M. and Kenny, D. A. (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51 (6), pp. 1173–82.

Barone, T. (2007) A return to the gold standard? Questioning the future of narrative construction as educational research. *Qualitative Inquiry*, 13 (4), pp. 454–70.

Barr Greenfield, T. (1975) Theory about organisations: a new perspective and its implications for schools. In M. G. Hughes (ed.) *Administering Education: International Challenge*. London: Athlone Press, pp. 71–99.

Barrett, M. S. and Mills, J. (2009) The inter-reflexive possibilities of dual observations: an account from and through experience. *International Journal of Qualitative Studies in Education*, 22 (4), pp. 417–29.

Barron, K. (1999) Ethics in qualitative social research on marginalized groups. *Scandinavian Journal of Disability Research*, 1 (1), pp. 38–49.

Bartgis, J. and Albright, G. (2016) Online role-play simulations with emotionally responsive avatars for the early detection of Native Youth Psychological Distress, including depression and suicidal ideation. *American Indian & Alaska Native Mental Health Research*, 23 (2), pp. 1–27.

Barthes, R. (1972) *Critical Essays*. Evanston, IL: Northwestern University Press.

Bartlett, J. E., II, Kotrlik, J. W. and Higgins, C. C. (2001) Organizational research: determining appropriate sample size in survey research. *Information Technology, Learning and Performance Journal*, 19 (1), pp. 43–50.

Bartlett, L. and Vavrus, F. (2016) *Rethinking Case Study Research: A Comparative Approach*. New York: Routledge.

Barton, A. (2002) Evaluation research as passive and apolitical? Some reflections from the field. *International Journal of Social Research Methodology*, 5 (4), pp. 371–8.

Barton, E. S., Walton, T. and Rowe, D. (1976) Using grid technique with the mentally handicapped. In P. Slater (ed.) *The Measurement of Intrapersonal Space by Grid Technique, Vol. 1: Explorations of Intrapersonal Space*. London: John Wiley & Sons, pp. 47–68.

Barton, K. C. (2015) Elicitation techniques: getting people to talk about ideas they don't usually talk about. *Theory and Research in Social Education*, 43 (2), pp. 179–205.

Bassett, R. and McGibbon, E. (2012) A critical participatory and collaborative method for scoping the literature. *Quality and Quantity*, 47 (6), pp. 3249–59.

Bassey, M. (1998) *Action Research for Improving Educational Practice*. In R. Halsall (ed.) *Teacher Research and School Improvement*. Buckingham, UK: Open University Press, pp. 167–78.

Bassey, M. (1999) *Case Study Research in Educational Settings*. Buckingham, UK: Open University Press.

Batliwala, S. and Patel, S. (2005) Enumeration. In R. Tandon (ed.) *Participatory Research: Revisiting the Roots*. New Delhi: Mosaic Books, pp. 295–312.

Batteson, C. and Ball, S. J. (1995) Autobiographies and interviews as means of 'access' to elite policy making in education. *British Journal of Educational Studies*, 43 (2), pp. 201–16.

Baudrillard, J. (2012) *The Ecstasy of Communication*. Los Angeles, CA: Semiotext(e).

Bauman, R. (1986) *Story, Performance and Event*. Cambridge: Cambridge University Press.

Baumgartner, M. (2008) Regularity theories reassessed. *Philosophia*, 36 (3), pp. 327–54.

Baumgartner, M. (2009) Inferring causal complexity. *Sociological Methods and Research*, 38 (1), pp. 71–101.

Baumrind, D. (1964) Some thoughts on ethics of research after reading Milgram's behavioral study of obedience. *American Psychologist*, 19 (6), pp. 421–3.

Bazeley, P. (2006) The contribution of computer software to integrating qualitative and quantitative data and analysis. *Research in the Schools*, 13 (1), pp. 64–74.

Bazeley, P. and Jackson, K. (eds) (2013) *Qualitative Data Analysis with NVivo* (second edition). London: Sage.

Beatty, P. C. and Willis, G. B. (2007) Research synthesis: the practice of cognitive interviewing. *Public Opinion Quarterly*, 71 (2), pp. 287–311.

Beck, R. N. (1979) *Handbook in Social Philosophy*. New York: Macmillan.

Becker, H. S. (1958) Problems of inference and proof in participant observation. *American Sociological Review*, 23 (6), pp. 652–60.

Becker, H. S. (1967) Whose side are we on? *Social Problems*, 14 (3), pp. 239–47.

Becker, H. S. (1986) *Doing Things Together: Selected Papers*. Evanston, IL: Northwestern University Press.

Becker, H. S. (1970) *Sociological Work*. Chicago, IL: Aldane.

Becker, H. S. (1998) *Tricks of the Trade: How to Think about Your Research while You're Doing It*. Chicago, IL: University of Chicago Press.

Becker, H. S. and Geer, B. (1960) Participant observation: the analysis of qualitative field data. In R. Adams and J. Preiss (eds) *Human Organization Research: Field Relations and Techniques*. Homewood, IL: Dorsey, pp. 267–89.

Beckett, C. and Clegg, S. (2007) Qualitative data from a postal questionnaire: questioning the presumption of the value of presence. *International Journal of Social Research Methodology*, 10 (4), pp. 307–17.

Belbase, S., Luitel, B. C. and Taylor, P. C. (2008) Autoethnography: a method of research and teaching for transformative education. *Journal of Education and Research*, 1 (1), pp. 86–95.

Bell, J. (1991) *Doing Your Research Project* (second edition). Milton Keynes, UK: Open University Press.

Bell, R. C. (2000) On testing the commonality of constructs in supplied grids. *Journal of Constructivist Psychology*, 13 (4), pp. 303–11.

Bell, R. C. (2004a) Predictive relationships in repertory grid data: a new elaboration of Kelly's organization corollary. *Journal of Constructivist Psychology*, 17 (4), pp. 281–95.

Bell, R. C. (2004b) A new approach to measuring conflict or inconsistency in grids. *Personal Construct Theory & Practice*, 1 (1), pp. 53–9.

Bell, R. C., Vince, J. and Costigan, J. (2002) Which vary more in repertory grid data: constructs or elements? *Journal of Constructivist Psychology*, 15 (4), pp. 305–14.

Belson, W. A. (1975) *Juvenile Theft: Causal Factors*. London: Harper & Row.

Belson, W. A. (1986) *Validity in Survey Research*. Aldershot: Gower.

Beneito-Montagut, R. (2017) Big data and educational research. In D. Wyse, E. Smith, L. E. Suter and N. Selwyn (eds) *The BERA/Sage Handbook of Educational Research*. London: Sage, pp. 913–33.

Beney, T. (2011) *Distinguishing Evaluation from Research*. Available from: www.uniteforsight.org/evaluation-course/module10 [Accessed 4 February 2016].

Bennett, L. and Nair, C. S. (2010) A recipe for effective participation rates for web-based surveys. *Assessment and Evaluation in Higher Education*, 35 (4), pp. 357–65.

Berg, J. (1957) Review of 'Supervisory and Executive Development. A Manual for Role Playing'. *Journal of Counseling Psychology*, 4 (4), pp. 332–3.

Berger, J. (1972) *Ways of Seeing*. London: British Broadcasting Corporation; Harmondsworth: Penguin.

Berger, P. L. and Luckmann, T. (1967) *The Social Construction of Reality*. Harmondsworth: Penguin.

Berger, R. (2015) Now I see it, now I don't: researcher's position and reflexivity in qualitative research. *Qualitative Research*, 15 (2), pp. 219–34.

Bergman, M. M. (2011a) The politics, fashions and conventions of research methods. *Journal of Mixed Methods Research*, 5 (2), pp. 99–102.

Bergman, M. M. (2011b) The good, the bad, and the ugly in mixed methods research and design. *Journal of Mixed Methods Research*, 5 (4), pp. 271–5.

Berg-Schlosser, D. (2012) *Mixed Methods in Comparative Politics: Principles and Applications*. Basingstoke, UK: Palgrave Macmillan.

Bernard, H. R. (1994) *Research Methods in Anthropology: Qualitative and Quantitative Approaches* (second edition). Walnut Creek, CA: AltaMira Press.

Bernstein, B. (1970) Education cannot compensate for society. *New Society*, February, 387, pp. 344–57.

Bernstein, B. (1974) Sociology and the sociology of education: a brief account. In J. Rex (ed.) *Approaches to Sociology: An Introduction to Major Trends in British Sociology*. London: Routledge & Kegan Paul, pp. 145–59.

Bernstein, B. (1975) Class and pedagogies: visible and invisible. In *Class, Codes and Control, Vol. 3*. London: Routledge & Kegan Paul, pp. 116–56.

Bernstein, R. J. (1983) *Beyond Objectivism and Relativism*. Oxford: Blackwell.

Best, D. (1992) *The Rationality of Feeling: Understanding the Arts in Education*. London: Falmer Press.

Bettez, S. C. (2015) Navigating the complexity of qualitative research in postmodern contexts: assemblage, critical reflexivity, and communion as guides. *International Journal of Qualitative Studies in Education*, 28 (8), pp. 932–54.

Beynon, H. (1988) Regulating research: politics and decision making in industrial organizations. In A. Bryman (ed.) *Doing Research in Organizations*. London: Routledge, pp. 21–33.

Bezzi, A. (1999) What is this thing called Geoscience? Epistemological dimensions elicited with the repertory grid and their implications for scientific literacy. *Science Education*, 83 (6), pp. 675–700.

Bhabha, H. (1994) *The Location of Culture*. London: Routledge.

Bhabha, H. (2004) *The Location of Culture* (second edition). London: Routledge.

Bhatti, G. (2012) Ethnographic and representational styles. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 80–4.

Bickel, R. (2007) *Multilevel Analysis for Applied Research*. New York: Guilford Press.

Biddle, B. J. and Anderson, D. S. (1991) Social research and educational change. In D. S. Anderson and B. J. Biddle (eds) *Knowledge for Policy: Improving Education through Research*. London: Falmer, pp. 1–20.

Bieri, J. (1955) Cognitive complexity-simplicity and predictive behavior. *Journal of Abnormal and Social Psychology*, 51 (2), pp. 263–8.

Biesta, G. (2007) Why 'What Works' won't work: evidence-based practice and the democratic deficit in educational research. *Educational Theory*, 57 (1), pp. 1–22.

Biesta, G. (2010a) Pragmatism and the philosophical foundations of mixed methods research. In A. Tashakkori and C. Teddlie (eds) *The Sage Handbook of Mixed Methods in Social and Behavioral Research* (second edition). Thousand Oaks, CA: Sage, pp. 95–115.

Biesta, G. (2010b) Why 'What Works' still won't work: from evidence-based education to value-based education. *Studies in Philosophy and Education*, 29 (5), pp. 491–503.

Biesta, G. (2012) Mixed methods. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 147–52.

Biesta, G., Allan, J. and Edwards, R. (2011) The theory question in research capacity building in education: towards an agenda for research and practice. *British Journal of Educational Studies*, 59 (3), pp. 225–39.

Billings, D. M. and Halstead, J. A. (2009) *Teaching in Nursing: A Guide for Faculty* (third edition). St. Louis, MO: Elsevier.

Binet, A. (1905) Méthode nouvelle pour le diagnostic de l'intelligence des anormaux. Cited in G. de Landsheere (1997) History of educational research. In J. P. Keeves (ed.) *Educational Research, Methodology, and Measurement: An International Handbook* (second edition). Oxford: Elsevier Science Ltd, pp. 8–16.

Birks, M. and Mills, J. (2015) *Grounded Theory: A Practical Guide* (second edition). London: Sage.

Birnbaum, M. H. (2009) Designing online experiments. In A. Joinson, K. McKenna, T. Postmes and U.-D. Reips (eds) *The Oxford Handbook of Internet Psychology*. Oxford: Oxford University Press, pp. 391–403.

Biziouras, N. (2013) Midshipmen form a coalition government in Belgium: lessons from a role-playing simulation. *Political Science and Politics*, 46 (2), pp. 251–6.

Black, P. (1998) *Testing: Friend or Foe?* London: Falmer.

Black, T. R. (1999) *Doing Quantitative in the Social Sciences*. London: Sage.

Blalock, H. M. (1979) *Social Statistics* (second edition). New York: McGraw-Hill.

Blalock, H. M. (1991) Dilemmas of social research. In D. S. Anderson and B. J. Biddle (eds) *Knowledge for Policy: Improving Education through Research*. London: Falmer, pp. 60–9.

Bland, M. (2010) The analysis of cluster-randomised trials in education. *Effective Education*, 2 (2), pp. 165–80.

Bless, H., Bohner, G., Traudel, H. and Schwartz, N. (1992) Asking difficult questions: task complexity increases the impact of impact alternatives. *European Journal of Social Psychology*, 22, pp. 309–12.

Blikstad-Balas, M. (2016) Key challenges of using video when investigating social practices in education: contextualization, magnification and representation. *International Journal of Research and Method in Education*. Available from: http://dx.doi.org/10.1080/1743727X.2016.1181162 [Accessed 10 October 2016].

Blix, S. B. and Wettergren, Å. (2015) The emotional labour of gaining and maintaining access to the field. *Qualitative Research*, 15 (6), pp. 688–704.

Bloom, B. (ed.) (1956) *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain*. London: Longman.

Blumenfeld-Jones, D. (1995) Fidelity as a criterion for practising and evaluating narrative inquiry. *International Journal of Qualitative Studies in Education*, 8 (1), pp. 25–33.

Blumer, H. (1969) *Symbolic Interactionism: Perspective and Method*. Englewood Cliffs, NJ: Prentice-Hall.

Boal, A. (1979) *Theatre of the Oppressed*. London: Pluto Press.

Boal, A. (2002) *Games for Actors and Non-Actors*. London: Routledge.

Boas, F. (1943) Recent anthropology. *Science*, 98, pp. 311–14.

Bochner, A. P. (2001) Narrative's virtues. *Qualitative Inquiry*, 7 (2), pp. 131–57.

Bochner, A. P. (2007) Notes towards an ethic of memory in autoethnographic inquiry. In N. K. Denzin and M. D. Giardina (eds) *Ethical Futures in Educational Research*. Walnut Creek, CA: West Coast Books, pp. 197–208.

Boellstorff, T. (2008) *Coming of Age in Second Life*. Princeton, NJ: Princeton University Press.

Boellstorff, T. (2015) *Coming of Age in Second Life: An Anthropologist Explores the Virtually Human*. Princeton, NJ: Princeton University Press.

Boellstorff, T., Nardi, B. and Taylor, T. L. (2012) *Ethnography and Virtual Worlds*: *A Handbook of Method*. Princeton, NJ: Princeton University Press.

Bogdan, R. G. and Biklen, S. K. (1992) *Qualitative Research for Education* (second edition). Boston, MA: Allyn & Bacon.

Bolton, G. (1996) Drama as research. In P. Taylor (ed.) *Researching Drama and Arts Education: Paradigms and Possibilities*. London: Falmer, pp. 187–94.

Bolton, G. and Heathcote, D. (1999) *So You Want to Use Role-Play? A New Approach in How to Plan*. Stoke-on-Trent, UK: Trentham Books Ltd.

Borenstein, M. (2009) Effect sizes for continuous data. In H. M. Cooper, L. V. Hedges and J. C. Valentine (eds) *The Handbook of Research Synthesis and Meta-analysis* (second edition). New York: The Russell Sage Foundation, pp. 221–35.

Borg, W. R. (1963) *Educational Research: An Introduction*. London: Longman.

Borg, W. R. and Gall, M. D. (1979) *Educational Research: An Introduction* (third edition). London: Longman.

Borg, W. R. and Gall, M. D. (1996) *Educational Research: An Introduction* (sixth edition). New York: Longman.

Borgatta, E. F. (1957) Supervisory and executive development: A manual for role-playing – a review. *American Sociological Review*, 22 (4), p. 477.

Borge, L. E. (2012) Comments on Stephen Gorard: mixed methods research in education. In Research Council of Norway (ed.) *Mixed Methods in Educational Research: Report on the March Seminar, 2012*, p. 15. Available from: www.uv.uio.no/ils/personer/vit/kirstik/publikasjoner-pdf-filer/klette.-mixed-methods.pdf [Accessed 8 September 2015].

Borkowsky, F. T. (1970) The relationship of work quality in undergraduate music curricula to effectiveness in instrumental music teaching in the public schools. *Journal of Experimental Education*, 39 (1), pp. 14–19.

Borman, G. D., Hewes, G. M., Overman, L. T. and Brown, S. (2003) Comprehensive school reform and achievement: a meta-analysis. *Review of Educational Research*, 73 (2), pp. 125–230.

Borsboom, D., Mellenbergh, G. D. and van Heerden, J. (2004) The concept of validity. *Psychological Review*, 111 (4), pp. 1061–71.

Boruch, R. F. (1997) *Randomized Experiments for Planning and Evaluation*. Applied Social Research Methods Series, vol. 44. Thousand Oaks, CA: Sage.

Boruch, R. F. and Cecil, J. S. (1979) *Assuring the Confidentiality of Social Research Data*. Philadelphia, PA: University of Pennsylvania Press.

Boston, M. D. (2008) Using classroom artifacts as evidence of quality instruction in mathematics. Paper presented at the Annual Conference of the Association of American Colleges of Teacher Education, New Orleans. Available from: www.allacademic.com/meta/p_mla_apa_research_citation/2/0/7/4/6/pages207464/p207464-2.php [Accessed 20 May 2010].

Boudon, R. (1973) *Education, Opportunity and Social Inequality*. New York: John Wiley & Sons.

Bouguen, A. and Gurgand, M. (2012) *Randomized Controlled Experiments in Education*. EENEE Analytical Report no. 11 for the European Commission. Paris: European Commission.

Boulton, J. G., Allen, P. M. and Bowman, C. (2015) *Embracing Complexity: Strategic Perspectives for an Age of Turbulence*. Oxford: Oxford University Press.

Bourdieu, P. (1976) The school as a conservative force. In R. Dale, G. Esland and M. MacDonald (eds) *Schooling and Capitalism*. London: Routledge & Kegan Paul, pp. 110–17.

Bourdieu, P. and Darbel, A. with Schnapper, D. (1991) *The Love of Art: European Art Museums and Their Public*. Cambridge: Polity Press.

Bourne-Day, J. and Lee-Treweek, G. (2008) Interconnecting lives: examining privacy as a shared concern for the researched and researchers. In B. Jegatheesan (ed.) *Access: A Zone of Comprehension, and Inclusion*. London: Emerald Group Publishing Limited, pp. 29–61.

Bowe, R., Ball, S. J. and Gold, A. (1992) *Reforming Education and Changing Schools*. London: Routledge.

Boyd, D. M. and Ellison, N. B. (2007) Social network sites: definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13 (1), pp. 210–30.

Bracht, G. H. and Glass, G. V. (1968) The external validity of experiments. *American Educational Research Journal*, 4 (5), pp. 437–74.

Brackertz, S. (2007) Who is hard to reach and why? *ISR Working Paper*. Adelaide: Institute for Social Research, Swinburne University of Technology. Available from: www.sisr.net/publications/0701brackertz.pdf [Accessed 16 February 2010].

Bradburn, N. M. and Sudman, S. (1979) *Improving Interview Method and Questionnaire Design*. San Francisco, CA: Jossey-Bass.

Bradley, B. A. and Reinking, D. (2011) Enhancing research and practice in early childhood through formative and design experiments. *Early Child Development and Care*, 181 (3), pp. 305–19.

Bradshaw, M. and Hulquist, B. L. (2017) *Innovative Teaching Strategies in Nursing and Related Health Professions*. Burlington, MA: Jones and Bartlett Publishers, Inc.

Brannen, J. (2005) Mixing methods: the entry of qualitative and quantitative approaches into the research process. *International Journal of Social Research Methodology*, 8 (3), pp. 173–84.

Bransford, J. D. and Schwartz, D. L. (1999) Rethinking transfer: a simple proposal with multiple implications. *Review of Research in Education*, 24 (1), pp. 61–100.

Braund, M., Moodley, T., Ekron, C. and Ahmed, Z. (2015) Crossing the border: science teachers using role-play in grade 7. *African Journal of Research in Mathematics, Science and Technology Education*, 19 (2), pp. 107–17.

Bray, M. and Lykins, C. (2012) *Shadow Education: Private Supplementary Tutoring and Its Implications for Policy Makers in Asia.* Hong Kong: Comparative Education Research Centre in collaboration with Asian Development Bank.

Breakwell, G. M. (2000) Interviewing. In G. M. Breakwell, S. Hammond and C. Fife-Shaw (eds) *Research Methods in Psychology* (second edition). London: Sage, pp. 239–50.

Breakwell, G. M., Hammond, S. and Fife-Shaw, C. (eds) (1995) *Research Methods in Psychology* (first edition). London: Sage.

Breakwell, G. M., Hammond, S., Fife-Shaw, C. and Smith, J. A. (eds) (2006) *Research Methods in Psychology* (third edition). London: Sage.

Breen, R. and Goldthorpe, J. H. (1997) Explaining educational differentials: towards a formal rational action theory. *Rationality and Society*, 9 (3), pp. 275–305.

Breen, R. and Goldthorpe, J. H. (1999) Class inequality and meritocracy: a critique of Saunders and an alternative analysis. *British Journal of Sociology*, 50 (1), pp. 1–27.

Breen, R. and Goldthorpe, J. H. (2002) Merit, mobility and method: another reply to Saunders. *British Journal of Sociology*, 53 (4), pp. 575–82.

Brenner, M. (2006) Interviewing in educational research. In J. Green, G. Camilli and P. Elmore (eds) *Complementary Methods for Research in Education*. Washington, DC: American Educational Research Association/Erlbaum, pp. 357–70.

Brenner, M. and Marsh, P. (eds) (1978) *The Social Contexts of Method*. London: Croom Helm.

Brenner, M., Brown, J. and Canter, D. (1985) *The Research Interview*. London: Academic Press Inc.

Breznau, N. (2016) Secondary observer effects: idiosyncratic errors in small-N secondary data analysis. *International Journal of Social Research Methodology*, 19 (3), pp. 301–18.

Brislin, R. W. (1970) Back-translation for cross-cultural research. *Journal of Cross-cultural Psychology*, 1 (3), pp. 185–216.

British Educational Research Association (2004) *Revised Ethical Guidelines for Educational Research*. Southwell, UK: British Educational Research Association. Available from: www.bera.ac.uk/files/guidelines/ethica1.pdf [Accessed 17 April 2010].

British Educational Research Association (2011) *Ethical Guidelines for Educational Research*. London: British Educational Research Association.

British Psychological Society (2005) *Code of Conduct, Ethical Principles and Guidelines*. Available from: www.bps.org.uk/document-download-area/document-download$.cfm?file_uuid=6D0645CC-7E96-C67F-D75E2648E5580115andext=pdf [Accessed 20 May 2007].

British Psychological Society (2013) *Ethics Guidelines for Internet-Mediated Research*. Leicester, UK: British Psychological Society.

British Psychological Society (2014) *Code of Human Research Ethics*. Leicester, UK: British Psychological Society.

British Sociological Association (2002) *Statement of Ethical Practice*. Durham, UK: British Sociological Association. Available from: www.britsoc.co.uk/NR/rdonlyres/801B9A62-5CD3-4BC2-93E1-FF470FF10256/0/StatementofEthicalPractice.pdf [Accessed 17 April 2010].

British Sociological Association (2006) *Visual Sociology Group's Statement of Ethical Practice (2006)*. Available from: www.visualsociology.org.uk/about/ethical_statement.php [Accessed 2 July 2016].

Broadribb, S., Peachey, A., Carter, C. and Westrap, F. (2009) Using Second Life at the Open University: how the virtual world can facilitate learning for staff and students. In C. Wankel and J. Kingsley (eds) *Higher Education in Virtual Worlds: Teaching and Learning in Second Life*. Bingley, UK: Emerald, pp. 203–20.

Brock-Utne, B. (1996) Reliability and validity in qualitative research within education in Africa. *International Review of Education*, 42 (6), pp. 605–21.

Brooks, R., te Riele, K. and Maguire, M. (2014) *Ethics and Education Research*. London: Sage.

Brown, A. L. (1992) Design experiments: theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences*, 2 (2), pp. 141–78.

Brown, E. C., Low, S., Smith, B. H. and Haggerty, K. P. (2011) Outcomes from a school-randomized controlled trial of Steps to Respect: a bullying prevention program. *School Psychology Review*, 40 (3), pp. 423–43.

Brown, L. D. (2005a) People-centred development and participatory research. In R. Tandon (ed.) *Participatory Research: Revisiting the Roots*. New Delhi: Mosaic Books, pp. 90–9.

Brown, L. D. (2005b) Ambiguities in participatory research. In R. Tandon (ed.) *Participatory Research: Revisiting the Roots*. New Delhi: Mosaic Books, pp. 197–202.

Brown, L. D. and Tandon, R. (2005) Ideology and political economy in inquiry: participatory research. In R. Tandon (ed.) *Participatory Research: Revisiting the Roots*. New Delhi: Mosaic Books, pp. 54–66.

Browne, K. (2005) Snowball sampling: using social networks to research non-heterosexual women. *International Journal of Social Research Methodology*, 8 (1), pp. 47–60.

Bruckman, A. S. (2004) Introduction: opportunities and challenges in methodology and ethics. In M. D. Johns, S. S. Chen and G. J. Hall (eds) *Online Social Research: Methods, Issues and Ethics*. New York: Peter Lang, pp. 15–24.

Bruner, J. S. (1986) *Actual Minds, Possible Worlds*. Cambridge, MA: Harvard University Press.

Bruner, J. S. (2004) Life as narrative. *Social Research*, 71 (3), pp. 691–710.

Bryant, A. and Charmaz, K. (2007) *The Sage Handbook of Grounded Theory*. London: Sage.

Bryeson, D., Manicom, L. and Kassam, Y. (2005) The methodology of the participatory research approach. In R. Tandon (ed.) *Participatory Research: Revisiting the Roots*. New Delhi: Mosaic Books, pp. 179–96.

Bryman, A. (2007a) Barriers to integrating quantitative and qualitative research. *Journal of Mixed Methods Research*, 1 (1), pp. 8–22.

Bryman, A. (2007b) The research question in social research: what is its role? *International Journal of Social Research Methodology*, 19 (1), pp. 5–20.

Bryman, A. (2008) *Social Research Methods* (third edition). Oxford: Oxford University Press.

Bryman, A. and Cramer, D. (1990) *Quantitative Data Analysis for Social Scientists*. London: Routledge.

Buch, K. and Wetzel, D. K. (2001) Analysing and realigning organizational culture. *Leadership and Organizational Development Journal*, 22 (1), pp. 40–3.

Buchanan, E. A. and Ess, C. (2009) Internet research ethics and the Institutional Review Board: current practices and issues. *Computers and Society*, 39 (3), pp. 43–9.

Buchanan, E. A. and Zimmer, M. (2012) Internet research ethics. *Stanford Encyclopedia of Philosophy*. Available from: http://plato.stanford.edu/entries/ethics-internet-research [Accessed 4 April 2016].

Buckingham, J., Beaman, R. and Wheldall, K. (2012) A randomized controlled trial of a MultiLit small group intervention for older low progress readers. *Effective Education*, 4 (1), pp. 1–26.

Buckley, C. and Waring, M. (2009) The evolving nature of grounded theory: experiential reflections on the potential of the method for analysing children's attitudes towards physical activity. *International Journal of Social Research Methodology*, 12 (4), pp. 317–34.

Buhler, C. and Allen, M. (1972) *Introduction to Humanistic Psychology*. Monterey, CA: Brooks/Cole.

Bulmer, M. (ed.) (1982) *Social Research Ethics*. London and Basingstoke: Macmillan.

Burbules, N. C. (2016) How we use and are used by social media in education. *Educational Theory*, 66 (4), pp. 551–65.

Burgess, R. G. (ed.) (1989) *The Ethics of Educational Research*. Lewes, UK: Falmer Press.

Burgess, R. G. (1993a) Biting the hand that feeds you? Educational research for policy and practice. In R. G. Burgess (ed.) *Educational Research and Evaluation for Policy and Practice*. Lewes, UK: Falmer, pp. 1–18.

Burgess, R. G. (ed.) (1993b) *Educational Research and Evaluation for Policy and Practice*. Lewes, UK: Falmer.

Burke, P. (2001) *New Perspectives on Historical Writing*. University Park, PA: Pennsylvania State University Press.

Burman, E. and Parker, I. (1993) *Discourse Analytical Research*. London: Routledge.

Burrell, G. and Morgan, G. (1979) *Sociological Paradigms and Organizational Analysis*. London: Heinemann Educational Books.

Busher, H. and James, N. (2015) In pursuit of ethical research: studying hybrid communities using online and face-to-face communications. *Educational Research and Evaluation*, 21 (2), pp. 168–81.

Butt, T. (1995) What's wrong with laddering? *Changes*, 13, pp. 81–7.

Butt, T. (2008) *George Kelly: The Psychology of Personal Constructs*. London: Palgrave Macmillan.

Buxton, C. (1956) *College Teaching: A Psychologist's View*. New York: Harcourt Brace.

Byatt, A. S. (1990) *Possession.* London: Chatto & Windus.

Byatt, A. S. (2009) *The Children's Book*. London: Chatto & Windus.

Byrne, D. and Callaghan, G. (2014) *Complexity Theory and the Social Sciences*. London: Routledge.

Cabral, R. J. (1987) Role playing as group intervention. *Small Group Research*, 18 (4), pp. 470–82.

Cain, T. (2011) Teachers' classroom-based action research. *International Journal of Research and Method in Education*, 34 (1), pp. 3–16.

Calder, J. (1979) Introduction to applied sampling. *Research Methods in Education and the Social Sciences* (Block 3, Part 4, DE304). Milton Keynes, UK: Open University Press.

Caldwell Cook, H. (1917) *The Play Way*. London: Heinemann.

Callawaert, S. (1999) Philosophy of education, Frankfurt critical theory, and the sociology of Pierre Bourdieu. In T. Popkewitz and L. Fendler (eds) *Critical Theories in Education: Changing Terrains of Knowledge and Politics*. London: Routledge, pp. 117–44.

Camburn, M., Goldring, E., Sebastian, J., May, H. and Huff, J. (2015) An examination of the benefits, limitations and challenges of conducting randomized controlled experiments with principals. *Educational Administration Quarterly*, 1–34. DOI: 10.1177/0013161x15617808 [Accessed 2 July 2016].

Campbell Collaboration. (n.d.) The Campbell Collaboration: What helps? What harms? Based on what evidence? Available from: www.campbellcollaboration.org [Accessed 20 September 2011].

Campbell, D. T. (1975) Degrees of freedom and the case study. *Comparative Political Studies*, 8 (2), pp. 178–93. Cited in R. K. Yin (2009) *Case Study Research* (fourth edition). Thousand Oaks, CA: Sage, p. 5.

Campbell, D. T. and Fiske, D. W. (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56 (2), pp. 81–105.

Campbell, D. T. and Stanley, J. (1963) *Experimental and Quasi-Experimental Designs for Research on Teaching*. Boston, MA: Houghton Mifflin.

Campbell, J. (2002) A critical appraisal of participatory methods in development research. *International Journal of Social Research Methodology*, 5 (1), pp. 19–29.

Campbell, J. P., Daft, R. L. and Hulin, C. L. (1982) *What to Study: Generating and Developing Research Questions*. Beverly Hills, CA: Sage.

Campbell, M. J., Julious, S. A. and Altman, D. G. (1995) Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparison. *British Medical Journal*, 311, pp. 1145–8.

Cannell, C. F. and Kahn, R. L. (1968) Interviewing. In G. Lindzey and A. Aronson (eds) *The Handbook of Social Psychology, Vol. 2: Research Methods*. New York: Addison Wesley, pp. 526–95.

Caplan, A. (1982) On privacy and confidentiality in social science research. In T. Beauchamp, R. Faden, R. Wallace and L. Walters (eds) *Ethical Issues in Social Science Research*. Baltimore, MD: Johns Hopkins University Press, pp. 315–28.

Capra, F. (1996) *The Web of Life*. New York: Anchor Books.

Capra, F. and Luisi, P. L. (2014) *The Systems View of Life: A Unifying Vision*. Cambridge: Cambridge University Press.

Caracelli, V. and Greene, J. (1993) Data analysis strategies for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 15 (2), pp. 195–207.

Cardinal, B. J., Tuominen, K. J. and Rintala, P. (2003) Psychometric assessment of Finnish versions of exercise-related measures of transtheoretical model constructs. *International Journal of Behavioral Medicine*, 19 (1), pp. 31–43.

Carlson, J. F. and Geisinger, K. F. (2014) *The Nineteenth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Carmines, E. G. and Zeller, R. A. (1979) *Reliability and Validity in Assessment*. Beverly Hills, CA: Sage.

Carpenter, J. R. and Kenward, M. G. (2013) *Multiple Imputation and Its Application*. New York: John Wiley & Sons.

Carr, W. (2005) The role of theory in the professional development of an educational theorist. *Pedagogy, Culture and Society*, 13 (3), pp. 333–45.

Carr, W. (2006) Education without theory. *British Journal of Educational Studies*, 54 (2), pp. 136–59.

Carr, W. and Kemmis, S. (1986) *Becoming Critical*. Lewes, UK: Falmer.

Carroll, C., Booth, A. and Cooper, K. (2011) A worked example of 'best fit' framework synthesis: a systematic review of views concerning the taking of some potential chemopreventive agents. *BMC Medical Research Methodology*, 11 (1), p. 29.

Carspecken, P. F. (1996) *Critical Ethnography in Educational Research*. London: Routledge.

Carspecken, P. F. and Apple, M. (1992) Critical qualitative research: theory, methodology, and practice. In M. LeCompte, W. L. Millroy and J. Preissle (eds) *The Handbook of Qualitative Research in Education*. London: Academic Press Ltd, pp. 507–53.

Carte, L. and Torres, R. M. (2014) Role-playing: a feminist-geopolitical analysis of the everyday workings of the Mexican state. *Gender, Place & Culture*, 21 (10), pp. 1267–84.

Cartwright, D. (1991) Basic and applied social psychology. In D. S. Anderson and B. J. Biddle (eds) *Knowledge for Policy: Improving Education through Research*. London: Falmer, pp. 3–31.

Cartwright, N. and Hardie, J. (2012) *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford: Oxford University Press.

Carver, R. P. (1978) The case against significance testing. *Harvard Educational Review*, 48 (3), pp. 378–99.

Casey, A. (2013) 'Seeing the trees not just the wood': steps and not just journeys in teacher action research. *Educational Action Research*, 21 (2), pp. 147–63.

Cavan, S. (1977) Review of J. D. Douglas's (1976) 'Investigative Social Review: Individual and Team Field Research'. *The American Journal of Sociology*, 83 (3), pp. 809–11.

Centre for Reviews and Dissemination (CRD) (2009) Systematic reviews: CRD's guidance for undertaking reviews in health care. Available from: www.york.ac.uk/inst/crd/index_guidance.htm [Accessed 30 September 2011].

Chamberlain, M. (1975) *Fenwomen: A Portrait of Women in an English Village*. London: Virago.

Chambers, K. (2003) How often do you have sex: problem gambling as a sensitive issue. Paper presented at the Twelfth International Congress on Gambling and Risk Taking, Vancouver, BC.

Champagne, M. V. (2014) *The Survey Playbook: How to Create the Perfect Survey*. CreateSpace Independent Publishing Platform. Available from: www.createspace.com.

Chang, H. (2008) *Autoethnography as Method*. Walnut Creek, CA: Left Coast Press.

Charmaz, K. (2000) Grounded theory: objectivist and constructionist methods. In N. K. Denzin and Y. S. Lincoln (eds) *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage, pp. 509–35.

Charmaz, K. (2002) Qualitative interviewing and grounded theory analysis. In J. F. Gubrium and J. A. Holstein (eds) *Handbook of Interview Research: Context and Method*. Thousand Oaks, CA: Sage, pp. 675–94.

Charmaz, K. (2006) *Constructing Grounded Theory: A Practical Guide through Qualitative Data Analysis*. London: Sage.

Chatterji, M. (2004) Evidence on 'what works': an argument for Extended-Term Mixed-Method (ETMM) Evaluation Designs. *Educational Researcher*, 33 (9), pp. 3–13.

Chelinsky, E. and Mulhauser, F. (1993) Educational evaluations for the US Congress: some reflections on recent experience. In R. G. Burgess (ed.) *Educational Research and Evaluation for Policy and Practice*. London: Falmer, pp. 44–60.

Cheng, A., Auerbach, M., Hunt, E. A., Chang, T. P., Pusic, M., Nadkarni, V. and Kessler, D. (2014) Designing and conducting simulation-based research. *Pediatrics*, 133 (6), pp. 1091–101.

Chomsky, N. (1959) Review of Skinner's *Verbal Behaviour*. *Language*, 35 (1), pp. 26–58.

Chong, P. W. and Graham, L. J. (2013) The 'Russian doll' approach: developing nested case-studies to support international comparative research in education. *International Journal of Research and Method in Education*, 36 (1), pp. 23–32.

Christian, L. M., Parsons, N. L. and Dillman, D. A. (2009) Designing scalar questions for web surveys. *Sociological Methods and Research*, 37 (3), pp. 393–425.

Cicourel, A. V. (1964) *Method and Measurement in Sociology*. New York: The Free Press.

Clark, A. (2006) *Anonymising Research Data*. Working Paper 7/06 for ESRC National Centre for Research Methods. Manchester: ESRC National Centre for Research Methods. Available from: http://eprints.ncrm.ac.uk/480/1/0706_anonymising_research_data.pdf [Accessed 19 May 2010].

Clark, A., Holland, C., Katz, J. and Peace, S. (2009) Learning to see: lessons from a participatory observation research project in public spaces. *International Journal of Social Research Methodology*, 12 (4), pp. 345–60.

Clark, A., Prosser, J. and Wiles, R. (2010) Ethical issues in image-based research. *Arts and Health*, 2 (1), pp. 81–93.

Clarke, A. (2007) Grounded theory: critiques, debates and situational analysis. In W. Outhwaite and S. P. Turner (eds) *Handbook of Social Science Methodology*. Thousand Oaks, CA: Sage, pp. 423–42.

Clarke, F. (1940) *Education and Social Change: An English Interpretation*. London: Sheldon Press.

Clegg, S. (2005) Evidence-based practice in educational research: a critical realist critique of systematic review. *British Journal of Sociology of Education*, 26 (3), pp. 415–28.

Clifford, J. and Marcus, G. E. (eds) (1986) *Writing Culture: The Poetics and Politics of Ethnography*. Berkeley, CA: University of California Press.

Clifton, J. (2006) A conversation analytical approach to business communication: the case of leadership. *Journal of Business Communication*, 43 (3), pp. 202–19.

Clogg, C. C. and Haritou, A. (1997) The regression method of causal inference and a dilemma confronting this method. In V. R. McKim and S. P. Turner (eds) *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*. Notre Dame, IN: University of Notre Dame Press, pp. 83–112.

Cobb, P., Confrey, J., diSessa A., Lehrer, R. and Schauble, L. (2003) Design experiments in educational research. *Educational Researcher*, 32 (1), pp. 9–13.

Coch, D. (2007) Neuroimaging research with children: ethical issues and case scenarios. *Journal of Moral Education*, 36 (1), pp. 1–18.

Coe, R. (2000) What is an 'effect size'? CEM Centre, University of Durham. Available from: www.cemcentre.org/renderpage.asp?linkid=30325016 [Accessed 7 January 2005].

Coe, R., Fitz-Gibbon, C. T. and Tymms, P. (2000) Promoting evidence-based education: the role of practitioners. Roundtable paper presented at the British Educational Research Association, University of Cardiff, UK, 7–10 September.

Cohen, J. and Stewart, I. (1995) *The Collapse of Chaos*. Harmondsworth: Penguin.

Cohen, A. and Wollack, J. A. (2010) *Handbook on Test Development: Helpful Tips for Creating Reliable and Valid Classroom Tests*. Madison, WI: University of Wisconsin-Madison, Testing & Evaluation Services. Available from: http://testing.wisc.edu/Handbook%20on%20Test%20Construction.pdf [Accessed 25 April 2010].

Cohen, D. K. and Garet, M. S. (1991) Reforming educational policy with applied social research. In D. S. Anderson and B. J. Biddle (eds) *Knowledge for Policy: Improving Education through Research*. London: Falmer, pp. 123–40.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences* (second edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992) A power primer. *Psychological Bulletin*, 112 (1), pp. 1–8.

Cohen, J. (1994) The earth is round ($\rho<0.05$). *American Psychologist*, 49 (12), pp. 997–1003.

Cohen, L. and Holliday, M. (1979) *Statistics for Education and Physical Education*. London: Harper & Row.

Cohen, L. and Holliday, M. (1982) *Statistics for Social Scientists*. London: Harper & Row.

Cohen, L. and Holliday, M. (1996) *Practical Statistics for Students*. London: Paul Chapman Publishing.

Cohen, L., Manion, L. and Morrison, K. R. B. (2010) *A Guide to Teaching Practice* (revised fifth edition). London: Routledge.

Coleman, J. S. (1991) Social policy research and societal decision making. In D. S. Anderson and B. J. Biddle (eds) *Knowledge for Policy: Improving Education through Research*. London: Falmer, pp. 113–22.

Coleridge, S. T. (1817) *Biographia Literaria*, chapter 14. Available from: www.bl.uk/collection-items/biographia-literaria-by-samuel-taylor-coleridge [Accessed 3 October 2016].

Collier, J. (1957) Photography in anthropology: a report on two experiments. *American Anthropologist*, 59, pp. 843–59.

Collins, J. S. and Duguid, P. (1989) Situated cognition and the culture of learning. *Educational Researcher*, 18 (1), pp. 32–42.

Collins, K. M. T., Onwuegbuzie, A. J. and Johnson, R. B. (2012) Securing a place at the table: a review and extension of legitimation criteria for the conduct of mixed research. *American Behavioral Scientist*, 56 (6), pp. 849–65.

Collmann, J. and Matei, S. A. (eds) (2016) *Ethical Reasoning in Big Data: An Exploratory Analysis*. Geneva, Switzerland: Springer.

Colorado State University (2016) *Survey Research*. Avilable from: http://writing.colostate.edu/guides/guide.cfm?guideid=68 [Accessed 6 March 2016].

Connell, R. W., Ashenden, D. J., Kessler, S. and Doswett, G. W. (1996) Making the difference: schools, families and social division. Cited in B. Limerick, T. Burgess-Limerick and M. Grace (1996) The politics of interviewing: power relations and accepting the gift. *International Journal of Qualitative Studies in Education*, 9 (4), pp. 449–60.

Connelly, F. M. and Clandinin, D. J. (1999) Narrative inquiry. In J. P. Keeves and G. Lakomski (eds) *Issues in Educational Research*. Oxford: Elsevier Science Ltd, pp. 32–40.

Connolly, P. (2003) *Ethical Principles for Researching Vulnerable Groups*. Coleraine: University of Ulster. Available from: www.ofmdfmni.gov.uk/ethicalprinciples.pdf [Accessed 17 February 2010].

Conover, N. J. (1971) *Practical Nonparametric Statistics*. New York: John Wiley.

Conrad, F. G. and Blair, J. (2009) Sources of error in cognitive interviews. *Public Opinion Quarterly*, 73 (1), pp. 32–55.

Convery, I. and Cox, D. (2012) A review of research ethics in internet-based research. *Practitioner Research in Higher Education*, 6 (1), pp. 50–7.

Cook, T. D. (1991) Postpositivist criticisms, reform associations, and uncertainties about social research. In D. S. Anderson and B. J. Biddle (eds) *Knowledge for Policy: Improving Education through Research*. London: Falmer, pp. 43–59.

Cook, T. D. and Campbell, D. T. (1979) *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton-Mifflin.

Cook, T. D., Cooper, H. M., Cordray, D. S., Hartman, H., Hedges, L. V., Light, R. J., Louis, T. and Mosteller, F. (1992) *Meta-analysis for Explanation: A Casebook*. New York: The Russell Sage Foundation.

Cooke, R. A. and Lafferty, J. C. (1989) *The Organizational Culture Inventory*. Plymouth, MI: Human Synergistics International.

Cooley, C. H. (1902) *Human Nature and the Social Order*. New York: Charles Scribner.

Coomber, R. (1997) Using the Internet for survey research. *Sociological Research Online*, 2 (2). Available from: www.socresonline.org.uk/socresonline/2/2/2.html [Accessed 14 November 2000].

Cooper, B. (1998) Using Bernstein and Bourdieu to understand children's difficulties with 'realistic' mathematics testing: an exploratory study. *Qualitative Studies in Education*, 11 (4), pp. 511–32.

Cooper, B. and Dunne, M. (2000) *Assessing Children's Mathematical Knowledge Social Class, Sex and Problem-Solving*. Buckingham, UK: Open University Press.

Cooper, B. and Glaesser, J. (2011) Using case-based approaches to analyse large datasets: a comparison of Ragin's fsQCA and fuzzy cluster analysis. *International Journal of Social Research Methodology*, 14 (1), pp. 31–48.

Cooper, B. and Glaesser, J. (2012) Qualitative work and the testing and development of theory: lessons from a study combining cross-case and within-case analysis via Ragin's QCA. *Forum: Qualitative Social Research*, 13 (2), p. 4.

Cooper, B. and Glaesser, J. (2016a) Qualitative Comparative Analysis, necessary conditions, and limited diversity: some problematic consequences of Schneider and Wagemann's Enhanced Standard Analysis. *Field Methods*, 28 (3), pp. 300–15.

Cooper, B. and Glaesser, J. (2016b) Exploring the robustness of set theoretic findings from a large n fsQCA: an illustration from the sociology of education. *International Journal of Social Research Methodology*, 19 (4), pp. 445–59.

Cooper, B., Glaesser, J., Gomm, R. and Hammersley, M. (2012) *Challenging the Qualitative–Quantitative Divide: Explorations in Case-Focused Causal Analysis*. London: Continuum.

Cooper, D. C. and Schindler, P. S. (2001) *Business Research Methods* (seventh edition). New York: McGraw-Hill.

Cooper, H. M. and Hedges, L. V. (2009) Research synthesis as a scientific process. In H. M. Cooper, L. V. Hedges and J. C. Valentine (eds) *The Handbook of Research Synthesis and Meta-analysis* (second edition). New York: The Russell Sage Foundation, pp. 3–16.

Corbin, J. and Morse, J. M. (2003) The unstructured interactive interview: issues of reciprocity and risks when dealing with sensitive topics. *Qualitative Inquiry*, 9 (3), pp. 335–54.

Corbin, J. and Strauss, A. (1990) Grounded theory research: procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13 (1), pp. 3–21.

Corbin, J. and Strauss, A. (2008) *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (third edition). London: Sage.

Corbin, J. and Strauss, A. (2015) *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (fourth edition). London: Sage.

Corey, S. M. (1953) *Action Research to Improve School Practice*. New York: Teachers College, Columbia University.

Cormack, M. (1992) *Ideology*. London: Batsford Ltd.

Cortina, J. M. and Landis, R. S. (2011) The earth is *not* round. *Organizational Research Methods*, 14 (92), pp. 332–49.

Coser, L. A. and Rosenberg, B. (1969) *Sociological Theory: A Book of Readings* (third edition). New York: Macmillan.

Cothran, D. J., Kulinna, P. H., Banville, D., Choi E., Amade-Escot, A., MacPhail, A., Macdonald, D., Richard, J.-F.

Sarmento, P. and Kirk, D. (2005) A cross-cultural investigation of the use of teaching styles. *Research Quarterly for Exercise and Sport*, 76 (2), pp. 193–201.

Cowley, P. and Stuart, M. (2015) Whipping them in: role-playing party cohesion with a Chief Whip. *Journal of Political Science Education*, 11 (2), pp. 190–203.

Coyer, S. H. and Gallo, A. M. (2005) Secondary analysis of data. *Journal of Pediatric Care*, 19 (1), pp. 60–3.

Crawford, S. D., Coupler, M. P. and Lamias, M. J. (2001) Web-surveys: perceptions of burdens. *Social Science Computer Review*, 19 (2), pp. 146–62.

Cresswell, M. J. and Houston, J. G. (1991) Assessment of the national curriculum: some fundamental considerations. *Educational Review*, 43 (1), pp. 63–78.

Creswell, J. W. (1994) *Research Design: Qualitative and Quantitative Approaches*. Thousand Oaks, CA: Sage.

Creswell, J. W. (1998) *Qualitative Inquiry and Research Design: Choosing among the Five Traditions*. Thousand Oaks, CA: Sage.

Creswell, J. W. (2002) *Educational Research: Planning, Conducting and Evaluating Quantitative and Qualitative Research*. Upper Saddle River, NJ: Merrill Prentice-Hall.

Creswell, J. W. (2009) Mapping the field of mixed methods research. *Journal of Mixed Methods Research*, 3 (2), pp. 95–108.

Creswell, J. W. (2011) Controversies in mixed methods research. In N. K. Denzin and Y. S. Lincoln (eds) *Handbook of Qualitative Research* (fourth edition). Thousand Oaks, CA: Sage, pp. 269–84.

Creswell, J. W. (2013) *Research Design: Qualitative, Quantitative and Mixed Methods Approaches* (fourth edition). Thousand Oaks, CA: Sage.

Creswell, J. W. and Plano Clark, V. L. (2011) *Designing and Conducting Mixed Methods Research* (second edition). Thousand Oaks, CA: Sage.

Creswell, J. W. and Tashakkori, A. (2007) Differing perspectives on mixed methods research. *Journal of Mixed Methods Research*, 1 (4), pp. 303–8.

Cronbach, L. J. (1949) *Essential of Psychological Testing* (first edition). New York: Harper & Row.

Cronbach, L. J. (1970) *Essentials of Psychological Testing* (third edition). New York: Harper & Row.

Cronqvist, L. and Berg-Schlosser, D. (2009) Multi-Value QCA (mvQCA). In B. Rihoux and C. C. Ragin (eds) *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques*. London: Sage, pp. 69–86.

Crow, G., Wiles, R., Heath, S. and Charles, V. (2006) Research ethics and data quality: the implications of informed consent. *International Journal of Social Research Methodology*, 9 (2), pp. 83–95.

Crowley, C., Harré, R. and Tagg, C. (2002) Qualitative research and computing: methodological issues and practices in using QSR NVivo and NUD*IST. *International Journal of Social Research Methodology*, 5 (3), pp. 193–7.

Crowston, K., Allen, E. E. and Heckman, R. (2012) Using natural language processing technology for qualitative data analysis. *International Journal of Social Research Methodology*, 15 (6), pp. 523–43.

Croxford, L. (2006) *The Youth Cohort Surveys: How Good Is the Evidence? Special CES Briefing No. 38*. Edinburgh, UK: Centre for Educational Sociology, University of Edinburgh.

Crudge, S. E. and Johnson, F. C. (2007) Using the repertory grid and laddering technique to determine the user's evaluative model of search engines. *Journal of Documentation*, 63 (2), pp. 259–80.

Cuff, E. G. and Payne, G. C. F. (eds) (1979) *Perspectives in Sociology*. London: George Allen & Unwin.

Cumming, G. (2012) *Understanding the New Statistics: Effect Sizes, Confidence Intervals and Meta-analysis*. New York: Routledge.

Cummings, L. (1985) Qualitative research in the infant classroom: a personal account. In R. Burgess (ed.) *Issues in Educational Research: Qualitative Methods*. Lewes, UK: Falmer, pp. 216–50.

Cunningham, G. K. (1998) *Assessment in the Classroom*. London: Falmer.

Cunningham, P. and Gardner, P. (2004) *Becoming Teachers: Texts and Testimonies 1907–1950*. London: Woburn Press.

Curr, D. (1994) Role play. *British Medical Journal*, 308 (6930), p. 725.

Curriculum, Evaluation and Management Centre (2000) *A Culture of Evidence*. Available from: http://cem.dur.ac.uk/ebeuk/culture.htm [Accessed 21 May 2000].

Curtis, B. (1978) Introduction. In B. Curtis and W. Mays (eds) *Phenomenology and Education*. London: Methuen, pp. ix–xxvi.

Dale, A. (2006) Quality issues with survey research. *International Journal of Social Research Methodology*, 9 (2), pp. 143–58.

Dale, A., Arber, S. and Procter, M. (1998) *Doing Secondary Analysis*. London: Unwin Hyman.

Dalli, C. and Stephenson, A. (2010) Involving children in research in early childhood education settings: opening up the issues. In J. Loveridge (ed.) *Involving Children and Young People in Research in Educational Settings.* Wellington, New Zealand: University of Wellington, Jessie Hetherington Centre for Educational Research, pp. 11–46. Available from: www.educationcounts.govt.nz/__data/assets/pdf_file/0005/80708/957_Involving-CYP-02092010.pdf [Accessed 23 June 2012].

Danby, S. J., Ewing, L. and Thorpe, K. J. (2011) The novice researcher: interviewing young children. *Qualitative Inquiry*, 17 (1), pp. 74–84.

Dancey, C. P. and Reidy, J. (2011) *Statistics without Maths for Psychology* (fifth edition). Harlow, UK: Pearson.

Data Protection Act 1984. London: HMSO.

David, M. (2002) Problems of participation. *International Journal of Social Research Methodology*, 5 (1), pp. 11–17.

Davis, B. and Sumara, D. (2005) Challenging images of knowing: complexity science and educational research. *International Journal of Qualitative Studies in Education*, 18 (3), pp. 305–21.

Davis, D. (2014) *Imagining the Real: Towards a New Theory of Drama in Education*. London: University of London Institute of Education Press.

Davis, M. S. (1971) That's interesting! Towards a phenomenology of sociology and a sociology of phenomenology. *Philosophy of the Social Sciences*, 1 (2), pp. 309–44.

Day, C. and Sammons, P. (2008) Combining qualitative and quantitative methodologies in research on teachers' loves,

work, and effectiveness: from integration to synergy. *Educational Researcher*, 37 (6), pp. 330–42.

Day, K. J. (1985) Perspectives on privacy: a sociological analysis. Unpublished PhD thesis, University of Edinburgh. Quoted in R. M. Lee (1993) *Doing Research on Sensitive Topics*. London: Sage.

De Laine, M. (2000) *Fieldwork, Participation and Practice*. London: Sage.

De Lisle, J. (2011) The benefits and challenges of mixing methods and methodologies. *Caribbean Curriculum*, 18 (1), pp. 87–120.

De Vaus, D. (1999) *Research Design in Social Research*. London: Sage.

De Vaus, D. A. (2001) *Research Design in Social Research*. London: Sage.

Deem, R. (1994) Researching the locally powerful: a study of school governance. In G. Walford (ed.) *Researching the Powerful in Education*. London: UCL Press, pp. 151–71.

Delamont, S. (1976) *Interaction in the Classroom*. London: Methuen.

Delamont, S. (1981) All too familiar? A decade of classroom research. *Educational Analysis*, 3 (1), pp. 69–83.

Delamont, S. (1992) *Fieldwork in Educational Settings: Methods, Pitfalls and Perspectives*. London: Falmer.

Delamont, S. (2007) Arguments against auto-ethnography. Paper presented at the British Educational Research Association Annual Conference, Institute of Education, University of London, 5–8 September 2007. Available from: www.leeds.ac.uk/educol/documents/168227.htm [Accessed 18 April 2016].

Delamont, S. (2009) The only honest thing: autoethnography, reflexivity and small crises in fieldwork. *Ethnography and Education*, 4 (1), pp. 51–63.

DeMunck, V. C. and Sobo, E. (eds) (1998) *Using Methods in the Field: A Practical Introduction and Casebook.* Walnut Creek, CA: AltaMira Press.

Dennett, D. C. (1978) *Brainstorms: Philosophical Essays on Mind and Psychology*. Brighton, UK: Harvester Press.

Dennison, S. T. (2011) Interdisciplinary role play between social work and theatre students. *Journal of Teaching in Social Work*, 33 (4), pp. 415–30.

Denscombe, M. (1995) Explorations in group interviews: an evaluation of a reflexive and partisan approach. *British Educational Research Journal*, 21 (2), pp. 131–48.

Denscombe, M. (2008) Communities of practice: a research paradigm for the mixed methods approach. *Journal of Mixed Methods Research*, 2 (3), pp. 270–83.

Denscombe, M. (2009a) *Ground Rules for Social Research: Guidelines for Good Practice* (second edition). Milton Keynes, UK: Open University Press.

Denscombe, M. (2009b) Item non-response rates: a comparison of online and paper questionnaires. *International Journal of Social Research Methodology*, 12 (4), pp. 281–91.

Denscombe, M. (2014) *The Good Research Guide* (fourth edition). Maidenhead, UK: Open University Press.

Denshire, A. (2014) On auto-ethnography. *Current Sociology Review*, 62 (6), pp. 831–50.

Denzin, N. K. (1970) *The Research Act in Sociology: A Theoretical Introduction to Sociological Methods*. London: Butterworth.

Denzin, N. K. (1989) *The Research Act: A Theoretical Introduction to Sociological Methods* (third edition). Englewood Cliffs, NJ: Prentice-Hall.

Denzin, N. K. (1990) On understanding emotion: the interpretive-cultural agenda. In T. D. Kemper (ed.) *Research Agendas in the Sociology of Emotions*. New York: State University of New York Press, pp. 85–116.

Denzin, N. K. (1997) Triangulation in educational research. In J. P. Keeves (ed.) *Educational Research, Methodology and Measurement: An International Handbook* (second edition). Oxford: Elsevier Science Ltd, pp. 318–22.

Denzin, N. K. (1999) Cybertalk and the method of instances. In S. Jones (ed.) *Doing Internet Research: Critical Issues and Methods for Examining the Net*. Thousand Oaks, CA: Sage, pp. 107–26.

Denzin, N. K. (2004) Reading film: using photos and video as social science material. In U. Flick, E. von Kardoff and I. Steinke (eds) A *Companion to Qualitative Research*. London: Sage, pp. 234–47.

Denzin, N. K. (2006) Analytic autoethnography or déjà vu all over again. *Journal of Contemporary Ethnography*, 35 (4), pp. 419–28.

Denzin, N. K. (2008) The new paradigm dialog and qualitative inquiry. *International Journal of Qualitative Studies in Education*, 21 (4), pp. 315–25.

Denzin, N. K. (2012) Triangulation 2.0. *Journal of Mixed Methods Research*, 6 (2), pp. 80–8.

Denzin, N. K. and Lincoln, Y. S. (eds) (1994) *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage.

Department of Health, Education and Welfare (1971) *The Institutional Guide to D.H.E.W. Policy on Protecting Human Subjects*. Washington, DC: DHEW.

Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., Hall, R., Koschmann, T., Lemke, J. L. and Sherin, M. G. (2010) Conducting video research in the learning sciences: guidance on selection, analysis, technology, and ethics. *The Journal of the Learning Sciences*, 19 (1), pp. 3–53.

Design-Based Research Collective (2003) Design-based research: an emerging paradigm for educational inquiry. *Educational Researcher*, 32 (1), pp. 5–8.

Deutskens, E., De Ruyter, K. and Wetzels, M. (2005) An assessment of measurement invariance between online and mail surveys. *Research Memoranda 3*. Maastricht: University of Maastricht, Faculty of Economics and Business Administration.

Deyle, D. L., Hess, G. and LeCompte, M. L. (1992) Approaching ethical issues for qualitative researchers in education. In M. LeCompte, W. L. Millroy and J. Preissle (eds) *The Handbook of Qualitative Research in Education*. London: Academic Press Ltd, pp. 597–642.

Diaz de Rada, V. (2005) Influence of questionnaire design on response to mail surveys. *International Journal of Social Research Methodology*, 8 (1), pp. 61–78.

Diaz de Rada, V. and Dominguez, J. A. (2015) The quality of responses to grid questions as used in Web questionnaires (compared with paper questionnaires). *International Journal of Social Research Methodology*, 18 (4), pp. 337–48.

Dicker, R. and Gilbert, J. (1988) The role of the telephone in educational research. *British Educational Research Journal*, 14 (1), pp. 65–72.

Dickson-Swift, V., James, E. L., Kippen, S. and Liamputtong, P. (2006) Blurring boundaries in qualitative health research on sensitive topics. *Qualitative Health Research*, 16 (6), pp. 853–71.

Dickson-Swift, V., James, E. L., Kippen, S. and Liamputtong, P. (2007) Doing sensitive research: what challenges do qualitative researchers face? *Qualitative Research*, 7 (3), pp. 327–53.

Dickson-Swift, V., James, E. L., Kippen, S. and Liamputtong, P. (2008) Risk to researchers in qualitative research on sensitive topics: issues and strategies. *Qualitative Health Research*, 18 (1), pp. 133–44.

Dickson-Swift, V., James, E. L., Kippen, S. and Liamputtong, P. (2009) Researching sensitive topics: qualitative research as emotion work. *Qualitative Research*, 9 (1), pp. 61–79.

Diener, E. and Crandall, R. (1978) *Ethics in Social and Behavioral Research*. Chicago, IL: University of Chicago Press.

Dietz, S. M. (1977) An analysis of programming DRL schedules in educational settings. *Behaviour Research and Therapy*, 15, pp. 103–11.

Dillman, D. A. (2007) *Mail and Internet Surveys: The Tailored Design Method* (second edition). New York: John Wiley.

Dillman, D. A. and Bowker, D. K. (2000) The web questionnaire challenge to survey methodologists. In U.-D. Reips and M. Bosnjak (eds) *Dimensions of Internet Science*. Lengerich, Germany: Pabst Science Publishers, pp. 159–78. Available from: http://survey.sesrc.wsu.edu/dillman/zuma_paper_dillman_bowker.pdf [Accessed 26 February 2005].

Dillman, D. A., Carley-Baxter, L. and Jackson, A. (1999) Skip pattern compliance in three test forms: a theoretical and empirical evaluation. *SESRC Technical Report #99–01*. Pullman, WA: Washington State University, Social and Economic Sciences Research Center.

Dillman, D., Smyth, J. and Christian, L. (2009) *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. New York: Wiley.

Dillman, D. A., Smyth, J. D. and Christian, L. M. (2014) *Internet, Phone, Mail and Mixed-Mode Surveys: The Tailored Design Method* (fourth edition). Hoboken, NJ: John Wiley & Sons.

Dillman, D. A., Smyth, J. D., Christian, L. M. and Stern, M. J. (2003) Multiple answer questions in self-administered surveys: the use of check-all-that-apply and forced-choice question formats. Paper presented at the American Statistical Association. San Francisco, CA.

Dillman, D. A., Tortora, R. D. and Bowker, D. (1998a) Influence of plain vs. fancy design in response rates for web surveys. Proceedings of Survey Methods Section, annual meeting of the American Statistical Association, Dallas, Texas. Available from: http://survey.sesrc.wsu.edu/dillman.papers.htm [Accessed 8 February 2005].

Dillman, D. A., Tortora, R. D. and Bowker, D. (1998b) Principles for constructing web surveys. Available from: http://survey.sesrc.wsu.edu/dillman/papers/websurveyppr.pdf [Accessed 8 February 2005].

Dillon, J. T. (1984) The classification of research questions. *Review of Educational Research*, 54 (3), pp. 327–61.

Dixon, N. F. (1987) *Our Own Worst Enemy*. London: Jonathan Cape.

Dixon-Woods, M. (2011) Using framework-based synthesis for conducting reviews of qualitative studies. *BMC Medicine*, 9 (3), pp. 39–40.

Dixon-Woods, M., Cavers, D., Agarwal, S., Annandale, E., Arthur, A., Harvey, J., Hsu, R., Katbamna, S., Olsen, R., Smith, L., Riley, R. and Sutton, A. J. (2006) Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Medical Research Methodology*, 6 (35), pp. 1–13.

Dixon-Woods, M., Fitzpatrick, R. and Roberts, K. (2001) Including qualitative research in systematic reviews: opportunities and problems. *Journal of Evaluation in Clinical Practice*, 7 (2), pp. 125–33.

Dobbert, M. L. and Kurth-Schai, R. (1992) Systematic ethnography: toward an evolutionary science of education and culture. In M. LeCompte, W. L. Millroy and J. Preissle (eds) *The Handbook of Qualitative Research in Education*. London: Academic Press Ltd, pp. 93–160.

Dochartaigh, N. O. (2002) *The Internet Research Handbook*. London: Sage.

Docherty, S. and Sandelowski, M. (1999) Focus on qualitative methods: interviewing children. *Research in Nursing and Health*, 22 (2), pp. 177–85.

Doll, W. E. (1993) *A Post-modern Perspective on Curriculum*. New York: Teachers College Press.

Donohoe, P. and O'Sullivan, C. (2015) The Bullying Prevention Pack: fostering vocabulary and knowledge on the topic of bullying and prevention with role-plays and discussions to reduce primary school bullying. *Scenario Journal for Drama and Theatre in Foreign and Second Language Education*, 9 (1), pp. 97–114.

Dooley, D. (2001) *Social Research Methods* (fourth edition). Upper Saddle River, NJ: Prentice-Hall.

Doron, I. (2007) Court of ethics: teaching ethics and ageing by means of role-playing. *Educational Gerontology*, 33 (9), pp. 737–58.

Douglas, H. (2004) The irreducible complexity of objectivity. *Synthese*, 138 (3), pp. 453–73.

Douglas, J. D. (1973) *Understanding Everyday Life*. London: Routledge & Kegan Paul.

Douglas, J. D. (1976a) *Investigative Social Research*. Beverly Hills, CA: Sage.

Douglas, J. W. B. (1976b) The use and abuse of national cohorts. In M. D. Shipman (ed.) *The Organization and Impact of Social* Research. London: Routledge & Kegan Paul, pp. 3–21.

Douglas Home, C. (2007) Revealing insight into the parallel world of gangs. *HeraldScotland*, 4 September. Available from: www.heraldscotland.com/revealing-insight-into-the-parallel-world-of-gangs-1.864507 [Accessed 2 June 2010].

Duckworth, A. L. and Yeager, D. S. (2015) Measurement matters: assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, 44 (4), pp. 237–51.

Duncan, M. G. (1968) *A Dictionary of Sociology*. London: Routledge & Kegan Paul.

Duncombe, J. and Jessop, J. (2002) 'Doing rapport' and the ethics of 'faking friendship'. In M. Mauthner, M. Birch, J. Jessop and T. Miller (eds) *Ethics in Qualitative Research*. London: Sage, pp. 107–22.

Dunne, C. (2011) The place of the literature review in grounded theory research. *International Journal of Social Research Methodology*, 14 (2), pp. 111–24.

Durkheim, E. (1982) *The Rules of Sociological Method*. Glencoe, IL: The Free Press.

Durrant, G. B. (2006) Missing data methods in official statistics in the United Kingdom: some recent developments. *Allgemeines Statistisches Archiv*, 90 (4), pp. 577–93.

Durrant, G. B. (2009) Imputation methods for handling non-response in practice: methodological issues and recent debates. *International Journal of Social Research Methodology*, 12 (4), pp. 293–304.

Duşa, A. (2016) *QCAGUI: Modern Functions for Qualitative Comparative Analysis*. R Package Version 2.2. Available from: http://cran.r-project.org/package=QCAGUI [Accessed 4 August 2016].

Dyer, C. (1995) *Beginning Research in Psychology*. Oxford: Blackwell.

Dyson, M. (2007) My story in a profession of stories: auto ethnography – an empowering methodology for educators. *Australian Journal of Teacher Education*, 32 (1), pp. 36–48.

Eady, S., Drew, V. and Smith, A. (2015) Doing action research: using communicative spaces to facilitate (transformative) professional learning. *Action Research*, 13 (2), pp. 105–22.

Eagleton, T. (1991) *Ideology*. London: Verso.

Eastabrooks, C. A., Field, P. A. and Morse, J. M. (1994) Aggregating qualitative findings: an approach to theory development. *Qualitative Health Research*, 4 (4), pp. 503–11.

Ebbinghaus, H. (1897) Über eine neue methode zur Prüfung geistiger Fähigkeiten. Cited in G. de Landsheere (1997) History of educational research. In J. P. Keeves (ed.) *Educational Research, Methodology, and Measurement: An International Handbook* (second edition). Oxford: Elsevier Science Ltd, pp. 8–16.

Ebbutt, D. (1985) Educational action research: some general concerns and specific quibbles. In R. Burgess (ed.) *Issues in Educational Research: Qualitative Methods*. Lewes, UK: Falmer, pp. 152–74.

Ebel, R. L. (1979) *Essentials of Educational Measurement* (third edition). Englewood Cliffs, NJ: Prentice-Hall.

Economic and Social Research Council (2015) *ESRC Framework for Research Ethics*. Swindon, UK: Economic and Social Research Council.

Economic and Social Research Council National Centre for Research Methods (2008) *Visual Ethics: Ethical Issues in Visual Research*. NCRM/011. Swindon, UK: Economic and Social Research Council.

Eder, D. and Fingerson, L. (2003) Interviewing children and adolescents. In J. A. Holstein and J. F. Gubrium (eds) *Inside Interviewing: New Lenses, New Concerns*. Thousand Oaks, CA: Sage, pp. 33–53.

Edwards, A. D. (1980) Patterns of power and authority in classroom talk. In P. Woods (ed.) *Teacher Strategies: Explorations in the Sociology of the School*. London: Croom Helm, pp. 237–53.

Edwards, R. and Holland, J. (2013) *What Is Qualitative Interviewing?* London: Sage.

Edwards, R. and Mauthner, M. (2002) Ethics and feminist research: theory and practice. In M. Mauthner, M. Birch, J. Jessop and T. Miller (eds) *Ethics in Qualitative Research*. London: Sage, pp. 14–31.

Eisenhart, M. (1998) On the subject of interpretive reviews. *Review of Educational Research*, 68 (4), pp. 391–9.

Eisenhart, M. (2001) Educational ethnography past, present, and future: ideas to think with. *Educational Researcher*, 30 (8), pp. 16–27.

Eisenhart, M. A. and Howe, K. R. (1992) Validity in educational research. In M. D. LeCompte, W. L. Millroy and J. Preissle (eds) *The Handbook of Qualitative Studies in Education*. New York: Academic Press, pp. 643–80.

Eisner, E. W. (1985) *The Art of Educational Evaluation*. Lewes, UK: Falmer.

Eisner, E. W. (1991) *The Enlightened Eye: Qualitative Inquiry and the Enhancement of Educational Practice*. New York: Macmillan.

Eisner, E. W. (1997) The promise and perils of alternative forms of data representation. *Educational Researcher*, 26 (6), pp. 4–10.

Eisner, E. W. (2008) Art and knowledge. In J. G. Knowles and A. L. Cole (eds) *Handbook of the Arts in Qualitative Research*. Thousand Oaks, CA: Sage, pp. 3–12.

Elbaz, F. (1990) Knowledge and discourse: the evolution of research into teachers' thinking. In C. Day, M. Pope and P. Denicola (eds) *Insights into Teachers' Thinking and Practice*. London: Falmer, pp. 15–42.

Elliot, D. L., Reid, K. and Baumfeld, V. (2016) Capturing visual metaphors and tales: innovative or elusive? *International Journal of Research and Method in Education*. Available from: http://dx.doi.org/10.1080/1743727X.2016.1181164 [Accessed 3 June 2016].

Elliott, J. (1978) What is action-research in schools? *Journal of Curriculum Studies*, 10 (4), pp. 355–7.

Elliott, J. (1991) *Action Research for Educational Change*. Buckingham, UK: Open University Press.

Elliott, J. (2005) Becoming critical: the failure to connect. *Educational Action Research*, 13 (4), pp. 359–73.

Ellis, C. (2004) *The Ethnographic, Vol. 1: A Methodological Novel about Autoethnography*. Walnut Creek, CA: AltaMira Press.

Ellis, C. and Bochner, A. P. (2000) Autoethnography, personal narrative, reflexivity. In N. K. Denzin and Y. S. Lincoln (eds) *Handbook of Qualitative Research* (second edition). Thousand Oaks, CA: Sage, pp. 733–68.

Ellis, C. and Bochner, A. P. (2006) Analysing analytic autoethnography. *Journal of Contemporary Ethnography*, 35 (4), pp. 429–49.

Ellis, C., Adams, T. E. and Bochner, A. P. (2011) Autoethnography: an overview. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 12 (1), pp. 1–12.

Ellis, P. D. (2010) *The Essential Guide to Effect Sizes*. Cambridge: Cambridge University Press.

Ely, M., Anzul, M., Friedman, T., Garner, D. and Steinmetz, A. (1991) *Doing Qualitative Research: Circles within Circles*. London: Falmer Press.

Emerald, E. and Carpenter, L. (2015) Vulnerability and emotions in research: risks, dilemmas and doubts. *Qualitative Inquiry*, 21 (8), pp. 741–50.

Enders, C. (2010) *Applied Missing Data Analysis*. New York: Guilford Press.

Engberg, M. E. (2004) Improving intergroup relations in higher education: a critical examination of the influence of educational interventions on racial bias. *Review of Educational Research*, 74 (4), p. 473.

Engeström, Y. (2011) From design experiments to formative interventions. *Theory & Psychology*, 21 (5), pp. 598–628.

English, H. B. and English, A. C. (1958) *A Comprehensive Dictionary of Psychological and Psychoanalytic Terms*. London: Longman.

EPPI-Centre (n.d.) *The Evidence for Policy and Practice Information*. Available from: eppi.ioe.ac.uk/cms [Accessed 3 July 2012].

Epting, F. R., Suchman, D. I. and Nickeson, K. J. (1971) An evaluation of elicitation procedures for personal constructs. *British Journal of Psychology*, 62 (4), pp. 513–17.

Ercikan, K. and Roth, W. M. (2006) What good is polarizing research into qualitative and quantitative? *Educational Researcher*, 35 (5), pp. 14–23.

Erickson, F. E. (1992) Ethnographic microanalysis of interaction. In M. LeCompte, W. L. Millroy and J. Preissle (eds) *The Handbook of Qualitative Research in Education*. London: Academic Press Ltd, pp. 201–26.

Errington, E. (1997) *Role Play*. HERDSA Green Guide No. 21. Canberra, Australia: Higher Education Research and Development Society of Australasia.

Ess, C. and the Association of Internet Researchers (2002) *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee*. Available from: www.aoir.org/reports/ethics.pdf [Accessed 14 February 2010].

Evans, C. (2014) Twitter for teaching: can social media be used to enhance the process of learning? *British Journal of Educational Technology*, 45 (5), pp. 902–15.

Evans, J. and Benefield, P. (2001) Systematic reviews of educational research: does the medical model fit? *British Educational Research Journal*, 27 (5), pp. 527–41.

Evans, J. R. and Mathur, A. (2005) The value of online surveys. *Internet Research*, 15 (2), pp. 195–219.

Evans, L. (2010) *What Is Virtual Ethnography?* Available from: www.inter-disciplinary.net/wp-content/uploads/2010/02/evanspaper.pdf [Accessed 23 April 2016].

Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011) *Cluster Analysis* (fifth edition). Chichester, UK: John Wiley & Sons.

Eysenbach, G. and Till, J. (2001) Ethical issues in qualitative research on internet communities. *British Medical Journal*, 323 (7321), pp. 1103–5.

Ezzy, D. (2002) *Qualitative Analysis: Practice and Innovation*. London: Routledge.

Fahie, D. (2014) Doing sensitive research sensitively: ethical and methodological issues in researching workplace bullying. *International Journal of Qualitative Methods*, 13 (1), pp. 19–36.

Fairclough, N. (1992) *Critical Language Awareness*. London: Longman.

Fairclough, N. (1995) *Critical Discourse Analysis: The Critical Study of Language*. London: Longman.

Fairclough, N. (2003) *Analysing Discourse: Textual Analysis for Social Research*. London: Routledge.

Falk, R. and Greenbaum, C. W. (1978) Significance tests die hard. *Theory and Psychology*, 5 (1), pp. 75–98.

Falk, R. and Greenbaum, C.W. (1995) Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theory & Psychology*, 5 (1), pp. 75–98.

Farberow, N. L. (ed.) (1963) *Taboo Topics*. New York: Atherton Press.

Farquharson, K. (2005) A different kind of snowball: identifying key policymakers. *International Journal of Social Research Methodology*, 8 (4), pp. 345–53.

Farrell, D. and Petersen, J. C. (2010) The growth of Internet research methods and the reluctant sociologist. *Sociological Inquiry*, 80 (1), pp. 114–25.

Farrimond, H. (2013) *Doing Ethical Research*. Basingstoke, UK: Palgrave Macmillan.

Faugier, J. and Sargeant, M. (1997) Sampling hard to reach populations. *Journal of Advanced Nursing*, 26 (4), pp. 790–7.

Fay, B. (1987) *Critical Social Science*. New York: Cornell University Press.

Feilzer, M. Y. (2010) Doing mixed methods research pragmatically: implications for the rediscovery of pragmatism as a research paradigm. *Journal of Mixed Methods Research*, 4 (1), pp. 6–16.

Feldt, L. S. and Brennan, R. L. (1993) Reliability. In R. Linn (ed.) *Educational Measurement*. New York: Macmillan, pp. 105–46.

Fendler, L. (1999) Making trouble: prediction, agency, critical intellectuals. In T. S. Popkewitz and L. Fendler (eds) *Critical Theories in Education: Changing Terrains of Knowledge and Politics*. London: Routledge, pp. 169–88.

Ferrance, E. (2000) *Action Research*. Providence, RI: Northeast and Islands Regional Educational Laboratory at Brown University. Available from: www.alliance.brown.edu/pubs/themes_ed/act_research.pdf [Accessed 16 April 2010].

Festinger, L. and Katz, D. (1966) *Research Methods in the Behavioral Sciences*. New York: Holt, Rinehart & Winston.

Fetscherin, M. and Lattemann, C. (2007) *User Acceptance of Virtual Worlds: An Explorative Study about Second Life.* Rollins College, FL/University of Potsdam, Germany.

Fetters, M. D. and Freshwater, D. (2015) The 1 + 1 = 3 integration challenge. *Journal of Mixed Methods Research*, 9 (2), pp. 115–17.

Feyerabend, P. (1975) *Against Method: Outline of an Anarchistic Theory of Knowledge*. London: New Left Books.

Field, A. (2000) *Cluster Analysis*. Available from: www.statisticshell.com/docs/cluster.pdf [Accessed 27 August 2016].

Fielding, N. (2004) Getting the most from archived qualitative data: epistemological, practical and professional obstacles. *International Journal of Social Research Methodology*, 7 (1), pp. 97–104.

Fielding, N. G. and Fielding, J. L. (1986) *Linking Data*. Beverly Hills, CA: Sage.

Fielding, N. G., Lee, R. M. and Blank, G. (eds) (2008) *The Sage Handbook of Internet Online Research Methods*. London: Sage.

Figueroa, P. (1998) The autobiographical account of the education of an African slave in eighteenth century England. In M. Erben (ed.) *Biography and Education*. London: Falmer Press, pp. 149–63.

Figueroa, S. K. (2008) The grounded theory analysis of audiovisual texts. *International Journal of Social Research Methodology*, 11 (1), pp. 1–12.

Finch, J. (1985) Social policy and education: problems and possibilities of using qualitative research. In R. G. Burgess (ed.) *Issues in Educational Research: Qualitative Methods*. Lewes, UK: Falmer, pp. 109–28.

Finch, J. (2004) Feminism and qualitative research. *International Journal of Social Research Methodology*, 7 (1), pp. 61–4.

Fine, G. A. and Sandstrom, K. L. (1988) *Knowing Children: Participant Observation with Minors*. Qualitative Research Methods Series 15. Thousand Oaks, CA: Sage.

Fine, M. (2010) *A Brief History of the Participatory Action Research Collective*. New York: Institute for Participatory Action Research and Design, City of New York Graduate Center. Available from: http://web.gc.cuny.edu/che/start.htm [Accessed 18 April 2010].

Finfgeld-Connett, D. (2014) Use of content analysis to conduct knowledge-building and theory-generating qualitative systematic reviews. *Qualitative Research*, 14 (3), pp. 341–52.

Finfgeld-Connett, D. and Johnson, E. D. (2012) Literature search strategies for conducting knowledge-building and theory-generating qualitative systematic reviews. *Journal of Advanced Nursing*. DOI: 10.1111/j.1365-2648.2012.06037.x.

Finkelstein, B. (1998) Revealing human agency: the uses of biography in the study of educational history. In C. Kridel (ed.) *Writing Educational Biography: Explorations in Qualitative Research*. New York: Garland Publishing, pp. 45–59.

Finlay-Johnson, H. (1912/2008) *The Dramatic Method of Teaching*. London: Kessinger Publishing.

Finn, C. E. (1991) What ails education research? In D. S. Anderson and B. J. Biddle (eds) *Knowledge for Policy: Improving Education through Research*. London: Falmer, pp. 39–42.

Finnegan, R. (1996) Using documents. In R. Sapsford and V. Jupp (eds) *Data Collection and Analysis*. London: Sage and the Open University Press, pp. 138–51.

Fisher, R. A. (1966) *The Design of Experiments* (eighth edition). New York: Haffner Publishing Company.

Fiske, J. (1995) Audiencing. In N. K. Denzin and Y. S. Lincoln (eds) *Handbook of Qualitative Methods*. London: Sage, pp. 188–98.

Fiske, S. T. (1993) Controlling other people: the impact of power on stereotyping. *American Psychologist*, 48 (6), pp. 621–8.

Fitz, J. and Halpin, D. (1994) Ministers and mandarins: educational research in elite settings. In G. Walford (ed.) *Researching the Powerful in Education*. London: UCL Press, pp. 32–50.

Fitz-Gibbon, C. T. (1996) *Monitoring Education: Indicators, Quality and Effectiveness*. London: Cassell.

Fitz-Gibbon, C. T. (1997) *The Value Added National Project: Final Report*. London: School Curriculum and Assessment Authority.

Flanagan, J. (1949) Critical requirements: a new approach to employee evaluation. Cited in E. C. Wragg (1994) *An Introduction to Classroom Observation*. London: Routledge.

Fleiss, J. L. and Berlin, J. A. (2009) Effect sizes for dichotomous data. In H. M. Cooper, L. V. Hedges and J. C. Valentine (eds) *The Handbook of Research Synthesis and Meta-analysis* (second edition). New York: The Russell Sage Foundation.

Flick, U. (1998) *An Introduction to Qualitative Research*. London: Sage.

Flick, U. (2004a) *An Introduction to Qualitative Research* (fourth edition). London: Sage.

Flick, U. (2004b) Design and process in qualitative research. In U. Flick, E. von Kardoff and I. Steinke (eds) *A Companion to Qualitative Research*. London: Sage, pp. 146–52.

Flick, U. (2009) *An Introduction to Qualitative Research* (fourth edition). London: Sage.

Flick, U., Garms-Homolová, V., Herrman, W. J., Kuck, J. and Röhnsch, G. (2012) 'I can't prescribe something just because someone asks for it…': using mixed methods in the framework of triangulation. *Journal of Mixed Methods Research*, 6 (2), pp. 97–110.

Flick, U., von Kardoff, E. and Steinke, I. (eds) (2004) *A Companion to Qualitative Research* (trans. B. Jenner). London: Sage.

Flyvbjerg, B. (2006) Five misunderstandings about case-study research. *Qualitative Inquiry*, 12 (2), pp. 219–45.

Fogelman, K. (2002) Surveys and sampling. In M. Coleman and A. R. J. Briggs (eds) *Research Methods in Educational Leadership*. London: Paul Chapman Publishing, pp. 93–108.

Foskett, N. H. and Hesketh, A. J. (1997) Constructing choice in continuous and parallel markets: institutional and school leavers' responses to the new post-16 marketplace. *Oxford Review of Education*, 23 (3), pp. 299–319.

Foster, P., Gomm, R. and Hammersley, M. (2000) Case studies as spurious evaluations: the example of research on educational inequalities. *British Journal of Educational Studies*, 48 (3), pp. 215–30.

Foucault, M. (1970) *The Order of Things*. London: Tavistock.

Foucault, M. (1990) *Discipline and Punish: The Birth of a Prison*. Harmondsworth, UK: Penguin.

Foucault, M. (1998) *The History of Sexuality: The Will to Knowledge*. London: Penguin.

Fowler, F. J., Jr (2009) *Survey Research Methods* (fourth edition). Thousand Oaks, CA: Sage.

Fowler, J., Cohen, L. and Jarvis, P. (2000) *Practical Statistics for Field Biology*. Chichester, UK: John Wiley & Sons.

Fox, D. J. (1969) *The Research Process in Education*. New York: Holt, Rinehart & Winston.

Fox, J., Murray, C. and Warm, A. (2003) Conducting research using web-based questionnaires: practical, methodological, and ethical considerations. *International Journal of Social Research Methodology*, 6 (2), pp. 167–80.

Francis, B. (2010) Gender, toys and learning. *Oxford Review of Education*, 36 (3), pp. 325–44.

Frankfort-Nachmias, C. and Nachmias, D. (1992) *Research Methods in the Social Sciences*. London: Edward Arnold.

Fransella, F. (1975) *Need to Change?* London: Methuen.

Fransella, F. (2003) *International Handbook of Personal Construct Psychology*. New York: John Wiley.

Fransella, F. and Bannister, D. (1977) *A Manual for Repertory Grid Technique*. London: Academic Press.

Fransella, F., Bell, R. and Bannister, D. (2004) *A Manual for the Repertory Grid Technique* (second edition). Chichester, UK: Wiley.

Fraser, H. (2004) Doing narrative research: analysing personal stories line by line. *Qualitative Social Work*, 3 (2), pp. 179–201.

Frazer, E. (2002) Citizenship and culture. In P. Dunleavy, A. Gamble, R. Heffernan, I. Holliday and G. Peele (eds) *Developments in British Politics, Vol. 6* (revised edition). Basingstoke, UK: Palgrave, pp. 203–18.

Fredericksen, J. R. and Collins, A. (1989) A systems approach to educational testing. *Educational Researcher*, 189, pp. 27–32.

Freedman, D. A. (1991) Statistical models and shoe leather. *Sociological Methodology*, 21 (2), pp. 291–313.

Freire, P. (1972) *Pedagogy of the Oppressed*. Harmondsworth: Penguin.

Freshwater, D. and Cahill, J. (2013) Paradigms lost and paradigms regained. *Journal of Mixed Methods Research*, 7 (1), pp. 3–5.

Freud, A. (1936) *The Ego and the Mechanisms of Defence*. New York: International Universities Press.

Frick, A., Bächtiger, M. T. and Reips, U.-D. (1999) Financial incentives, personal information and dropout rate in online studies. In U.-D. Reips and M. Bosnjak (eds) *Dimensions of Internet Science*. Lengerich, Germany: Pabst Science, pp. 209–19.

Fricker, R. D., Jr and Schonlau, M. (2002) Advantages and disadvantages of Internet research surveys: evidence from the literature. *Field Methods*, 14 (4), pp. 347–67.

Friedman, H. H. and Amoo, T. (1999) Rating the rating scales. *Journal of Marketing Management*, 9 (3), pp. 114–23.

Frisbie, D. (1981) The relative difficulty ratio: a test and item index. *Educational and Psychological Measurement*, 41 (2), pp. 333–9.

Frost, J. L., Wortham, S. C. and Reifel, S. (2008) *Play and Child Development* (third edition). Upper Saddle River, NJ: Pearson.

Frueh, F. W. (2009) Back to the future: why randomized controlled trials cannot be the answer to pharmacogenomics and personalized medicine. *Pharmacogenetics*, 10 (7), pp. 1077–81.

Furlong, J. and Oancea, A. (2005) *Assessing Quality in Applied and Practice-Based Educational Research*. Oxford: Department of Educational Studies, University of Oxford. Available from: www.esrc.ac.uk/ESRCInfoCentre/Images/assessing_quality_shortreport_tcm6-8232.pdf [Accessed 10 October 2008].

Gadamer, H. G. (1975) *Truth and Method*. New York: Polity Press.

Gadd, D. (2004) Making sense of interviewee–interviewer dynamics in narratives about violence in intimate relationships. *International Journal of Social Research Methodology*, 7 (5), pp. 383–401.

Gage, N. L. (1989) The paradigm wars and their aftermath. *Teachers College Record*, 91 (2), pp. 135–50.

Gallagher, D. J. (2004) Educational research, philosophical orthodoxy and unfulfilled promises: the quandary of traditional research in US special education. In G. Thomas and R. Pring (eds) *Evidence-Based Practice in Education*. Buckingham, UK: Open University Press, pp. 119–30.

Galpin, C. J. (1915) *The Social Anatomy of an Agricultural Community*. Bulletin no. 34. Madison, WI: University of Wisconsin Agricultural Experiment Station. Available from: https://babel.hathitrust.org/cgi/pt?id=osu.32435011794021;view=1up;seq=5 [Accessed 2 October 2016].

Galton, M. and Simon, B. (1980) *Inside the Primary Classroom*. London: Routledge.

Garcia, A. S., Morrison, K. R. B., Tsoi, A. C. and He, J. M. (2014) *Managing Complex Change in School: Engaging Pedagogy, Technology, Learning and Leadership.* London: Routledge.

García-Horta, J. B. and Guerra-Ramos, M. T. (2009) The use of CAQDAS in educational research: some advantages, limitations and potential risks. *International Journal of Research and Method in Education*, 32 (2), pp. 151–65.

Gardner, G. (1978) *Social Surveys for Social Planners*. Milton Keynes, UK: Open University Press.

Gardner, H. (1993) *Multiple Intelligences: The Theory in Practice*. New York: Basic Books.

Garfinkel, H. (1967) *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.

Garrahan, P. and Stewart, P. (1992) *The Nissan Enigma: Flexibility at Work in a Local Economy*. London: Mansell.

Gaskell, G. D., O'Muircheartaigh, C. A. and Wright, D. B. (1994) Survey questions about the frequency of vaguely defined events: the effects of response alternatives. *Public Opinion Quarterly*, 58 (2), pp. 241–54.

Gee, J. P. (1996) *Social Linguistics and Literacies: Ideology in Discourse*. London: Routledge.

Gee, J. P. (2005) *An Introduction to Discourse Analysis Theory and Method* (second edition). London: Routledge.

Geertz, C. (1973) *The Interpretation of Cultures.* New York: Basic Books.

Geertz, C. (1974) From the native's point of view: on the nature of anthropological understanding. *Bulletin of the American Academy of Arts and Sciences*, 28 (1), pp. 26–45.

George, A. L. and Bennett, A. (2005) *Case Studies and Theory Development in the Social Sciences.* Cambridge, MA: MIT Press.

Geuss, R. (1981) *The Idea of a Critical Theory*. London: Cambridge University Press.

Gewirtz, S. and Ozga, J. (1993) Sex, lies and audiotape: interviewing the education policy elite. Paper presented to the Economic and Research Council, 1988 Education Reform Act research seminar, University of Warwick.

Gewirtz, S. and Ozga, J. (1994) Interviewing the education policy elite. In G. Walford (ed.) *Researching the Powerful in Education*. London: UCL Press, pp. 186–203.

Gibbs, A. (2012) Focus groups and group interviews. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 186–92.

Gibbs, G. R. (2007) *Analysing Qualitative Data*. London: Sage.

Gibbs, G. R. (2012) Software and qualitative data analysis. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 251–8.

Gibbs, G. R., Lewins, A. and Silver, C. (2005) *What Software Does and Does Not Do*. Available from: http://onlineqda.hud.ac.uk/Intro_CAQDAS/What_the_sw_can_do.php [Accessed 9 June 2016].

Gibbs, P., Cartney, P., Wilkinson, K., Parkinson, J., Cunningham, S., Janes-Reynolds, A., Zoubir, T., Brown, V., Barter, P., Sumber, P., MacDonald, A., Dayananda, A. and Pitt, A. (2016) Literature review on the use of action research in higher education. *Educational Action Research*. Available from: http://dx.doi.org/10.1080/09650792.2015.1124046 [Accessed 9 July 2016].

Gibson, R. (1985) Critical times for action research. *Cambridge Journal of Education*, 15 (1), pp. 59–64.

Giddens, A. (1975) *Positivism and Sociology*. London: Heinemann.

Giddens, A. (1976) *New Rules of Sociological Method: A Positive Critique of Interpretative Sociologies*. London: Hutchinson.

Giddens, A. (1979) *Central Problems in Social Theory*. London: Macmillan.

Giddens, A. (1984) *The Constitution of Society*. Cambridge: Polity Press.

Giddings, L. S. (2006) Mixed methods research: positivism dressed in drag? *Journal of Research in Nursing*, 11 (3), pp. 195–203.

Giddings, L. S. and Grant, B. M. (2007) A Trojan horse for positivism? A critique of mixed methods research. *Advances in Nursing Research*, 30 (1), pp. 52–60.

Gillies, V. and Alldred, P. (2002) The ethics of intention: research as a political tool. In M. Mauthner, M. Birch, J. Jessop and T. Miller (eds) *Ethics in Qualitative Research*. London: Sage, pp. 32–52.

Ginsberg, G. P. (1978) Role playing and role performance in social psychological research. In M. Brenner and P. Marsh (eds) *The Social Context of Method*. London: Croom Helm, pp. 91–121.

Gipps, C. (1994) *Beyond Testing: Towards a Theory of Educational Assessment*. London: Falmer.

Giroux, H. A. (1983) *Theory and Resistance in Education*. London: Heinemann.

Giroux, H. A. (1989) *Schooling for Democracy*. London: Routledge.

Glaesser, J. (2015) *Young People's Educational Careers in England and Germany. Integrating Survey and Interview Analysis via Qualitative Comparative Analysis*. Basingstoke, UK: Palgrave Macmillan.

Glaesser, J. and Cooper, B. (2011) Selecting cases for in-depth study from a survey dataset: an application of Ragin's configurational methods. *Methodological Innovations Online*, 6 (2), pp. 52–70.

Glaser, B. G. (1963) Retreading research materials: the use of secondary data by the independent researcher. *The American Behavioral Scientist*, 6 (10), pp. 11–14.

Glaser, B. G. (1978) *Theoretical Sensitivity: Advances in the Methodology of Grounded Theory*. Mill Valley, CA: Sociology Press.

Glaser, B. G. (1992) *Basics of Grounded Theory Analysis*. Mill Valley, CA: Sociology Press.

Glaser, B. G. (1996) Grounded theory: an interview with Barney Glaser. Video material for Program 8 of the course 'Doing a PhD in Business and Management'. University of Stirling and Heriot-Watt University.

Glaser, B. G. (1998) *Doing Grounded Theory: Issues and Discussions*. Mill Valley, CA: Sociology Press.

Glaser, B. G. and Strauss, A. L. (1967) *The Discovery of Grounded Theory*. Chicago, IL: Aldane.

Gläser, J. and Laudel, G. (2013) Life with and without coding: two methods for early-stage data analysis in qualitative research aiming at causal explanations. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 14 (2), pp. 1–24.

Glaser, J., Dixit, J. and Green, D. (2002) Studying hate crime with the internet: what makes racists advocate racist violence? *Journal of Social issues*, 58 (1), pp. 177–93.

Glass, G. V. (1976) Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5 (10), pp. 3–8.

Glass, G. V. (2000) *Meta-analysis at 25*. Available from: http://glass.ed.asu.edu/gene/papers/meta25.htmlGlass [Accessed 10 October 2012].

Glass, G. V. (2006) Meta-analysis: the quantitative synthesis of research findings. In J. L. Green, G. Camilli and P. B. Elmore (eds) *Handbook of Complementary Methods in Education Research* (third edition). Washington, DC: American Educational Research Association, pp. 427–38.

Glass, G. V. and Hopkins, K. D. (1996) *Statistical Methods in Education and Psychology* (third edition). Boston: Allyn & Bacon.

Glass, G. V. and Worthen, B. R. (1971) Evaluation and research: similarities and differences. *Curriculum Theory Network*, 3 (Fall), pp. 149–65.

Glass, G. V., McGaw, B. and Smith, M. L. (1981) *Meta-analysis in Social Research*. Beverly Hills, CA: Sage.

Gleason, B. (2016) New literacies practices of teenage Twitter users. *Learning, Media and Technology*, 41 (1), pp. 31–54.

Glover, D. and Bush, T. (2005) The online or e-survey: a research approach for the ICT age. *International Journal of Research and Method in Education*, 28 (2), pp. 135–46.

Gobo, G. (2011) Ethnography. In D. Silverman (ed.) *Qualitative Research* (third edition). London: Sage, pp. 15–34.

Goedert, J. D. and Rokooei, S. (2016) Project-based construction education with simulations in a gaming environment. *International Journal of Construction Education and Research*, 12 (3), pp. 208–23.

Goffman, E. (1963) *Stigma: Notes on the Management of Spoiled Identity*. Englewood Cliffs, NJ: Prentice-Hall.

Goffman, E. (1968) *Asylums*. Harmondsworth: Penguin.

Goffman, E. (1969) *The Presentation of Self in Everyday Life*. Harmondsworth: Penguin.

Goffman, E. (1976) The presentation of self in everyday life. In J. E. Combs and M. W. Mansfield (eds) *Drama in Life: The Uses of Communication in Society*. New York: Hastings House Publishers, pp. 62–72.

Golafshani, N. (2003) Understanding reliability and validity in qualitative research. *The Qualitative Report*, 8 (4), pp. 597–607. Available from: www.nova.edu/ssss/QR/QR8-4/golafshani.pdf [Accessed 29 October 2005].

Goldacre, B. (2013) *Building Evidence into Education*. London: Department for Education.

Goldstein, H. I. (1987) *Multilevel Modelling in Educational and Social Research*. London: Charles Griffin and Co. Ltd.

Goldstein, H. I. (2003) *Multilevel Statistical Models* (third edition). London: Edward Arnold.

Goldthorpe, J. H. (2007) *On Sociology, Vol. 2: Illustration and Retrospect* (second edition). Stanford, CA: Stanford University Press.

Gonzales, L., Brown, M. S. and Slate, J. R. (2008) Teachers who left the teaching professions: a qualitative understanding. *The Qualitative Report*, 13 (1), pp. 1–11.

Good, C. V. (1963) *Introduction to Educational Research*. New York: Appleton-Century-Crofts.

Goodwin, B. (2000) Out of control into participation. *Emergence*, 2 (4), pp. 40–9.

Gorard, S. (2001a) A changing climate for educational research? The role of research capability-building. Paper presented at the British Educational Research Association annual conference, University of Leeds, September.

Gorard, S. (2001b) *Quantitative Methods in Educational Research: The Role of Numbers Made Easy*. London: Continuum.

Gorard, S. (2002) Fostering scepticism: the importance of warranting claims. *Evaluation and Research in Education*, 16 (3), pp. 136–49.

Gorard, S. (2003) *Quantitative Methods in Social Science*. London: Continuum.

Gorard, S. (2005) Academies as the 'future of schooling': is this an evidence-based policy? *Journal of Education Policy*, 20 (3), pp. 369–77.

Gorard, S. (2007) The dubious benefits of multi-level modelling. *International Journal of Research and Method in Education*, 30 (2), pp. 221–36.

Gorard, S. (2012) Mixed methods research in education: some challenges and problems. In Research Council of Norway (ed.) *Mixed Methods in Educational Research: Report on the March Seminar, 2012*, pp. 5–13. Available from: www.uv.uio.no/ils/personer/vit/kirstik/publikasjoner-pdf-filer/klette.-mixed-methods.pdf [Accessed 8 September 2015].

Gorard, S. (2013) *Research Design*: *Creating Robust Approaches for the Social Sciences*. London: Sage.

Gorard, S. (2014) The link between Academies in England, pupil outcomes and local patterns of socio-economic segregation between schools. *Research Papers in Education*, 29 (3), pp. 268–84.

Gorard, S. (2016) Damaging real lives through obstinacy: re-emphasising why significance testing is wrong. *Sociological Research Online*, 21 (1), p. 2. DOI: 10.5153/sro.3857.

Gorard, S. and Gorard, J. (2016) What to do instead of significance testing? Calculating the 'number of counterfactual cases needed to disturb a finding'. *International Journal of Social Research Methodology*, 19 (4), pp. 481–90.

Gorard, S. and Smith, E. (2006) Editorial: combining numbers with narratives. *Evaluation and Research in Education*, 19 (2), pp. 59–62.

Gorard, S. and Taylor, C. (2004) *Combining Methods in Educational and Social Research*. Buckingham, UK: Open University Press.

Gorard, S. and Torgerson, C. (2006) The ESRC Researcher Development Initiative: promise and pitfalls of pragmatic trials in education. Paper presented at the British Educational Research Association Annual Conference, University of Warwick, 6–9 September.

Gorard, S., Roberts, K. and Taylor, C. (2004) What kind of creature is a design experiment? *British Educational Research Journal*, 30 (4), pp. 575–88.

Gordon, S. and Thomas, I. (2016) 'The learning sticks': reflections on a case study of role-playing for sustainability. *Environmental Education Research*, 22 (1), pp. 1–19.

Gordon, T. and Lahelma, E. (2003) From ethnography to life history: tracing transitions of school students. *International Journal of Social Research Methodology*, 6 (3), pp. 245–54.

Gough, D., Oliver, S. and Thomas, J. (2012) Introducing systematic reviews. In D. Gough, S. Oliver and J. Thomas (eds) *An Introduction to Systematic Reviews*. London: Sage, pp. 1–16.

Graham, A., Powell, M., Taylor, N., Anderson, D. and Fitzgerald, R. (2013) *Ethical Research Involving Children*. Florence, Italy: UNICEF Office of Research–Innocenti.

Graue, M. E. and Walsh, D. J. (1998) *Studying Children in Context: Theories, Methods and Ethics*. London: Sage.

Greckhamer, T. and Koro-Ljungberg, M. (2005) The erosion of a method: examples from grounded theory. *International Journal of Qualitative Studies in Education*, 18 (6), pp. 729–50.

Greene, J. C. (2005) The generative potential of mixed methods inquiry. *International Journal of Research and Method in Education*, 28 (2), pp. 207–11.

Greene, J. C. (2008) Is mixed methods social inquiry a distinctive methodology? *Journal of Mixed Methods Research*, 2 (1), pp. 7–22.

Greenhalgh, T., Robert, G., Macfarlane, F., B. P., Kyriakidou, O. and Peacock, R. (2005) Storylines of research in diffusion of innovation: a meta-narrative approach to systematic review. *Social Science and Medicine*, 61 (2), pp. 417–30.

Greenhow, C. and Gleeson, B. (2012) Twitteracy: tweeting as a new literacy practice. *The Educational Forum*, 76 (4), pp. 464–78.

Gregory, S., Lee, M., Dalgarno, B. and Tynan, B. (eds) (2015) *Virtual Worlds for Online Learning: Cases and Applications*. Hauppauge, NY: Nova Science Publishers.

Greig, A. D. and Taylor, J. (1999) *Doing Research with Children*. London: Sage.

Gronlund, N. (1981) *Measurement and Evaluation in Teaching* (fourth edition). New York: Collier-Macmillan.

Gronlund, N. E. (1985) *Stating Objectives for Classroom Instruction* (third edition). New York: Macmillan.

Gronlund, N. E. and Brookhart, S. M. (2008) *Gronlund's Writing Instructional Objectives* (eighth edition). New York: Pearson.

Gronlund, N. E. and Linn, R. L. (1990) *Measurement and Evaluation in Teaching* (sixth edition). New York: Macmillan.

Grosvenor, I., Lawn, M. and Rousmaniere, K. (eds) (1999) *Silences and Images: The Social History of the Classroom*. New York: Peter Lang.

Grumet, M. (1998) Research conversations: visible pedagogies, generous pedagogies. In J. Saxton and C. Millers (eds) *The Research of Practice: The Practice of Research*. Victoria, BC: International Drama in Education Research Institute, pp. 1–17.

Grundy, S. (1987) *Curriculum: Product or Praxis*. Lewes, UK: Falmer.

Grundy, S. (1996) Towards empowering leadership: the importance of imagining. In O. Zuber-Skerritt (ed.) *New Directions in Action Research*. London: Falmer, pp. 106–20.

Grundy, S. and Kemmis, S. (1988) Educational action research in Australia: the state of the art (an overview). In S. Kemmis and R. McTaggart (eds) *The Action Research*

Reader (second edition). Geelong, Victoria: Deakin University Press, pp. 83–97.
Guba, E. G. and Lincoln, Y. S. (1989) *Fourth Generation Evaluation*. Beverly Hills, CA: Sage.
Guba, E. G. and Lincoln, Y. S. (1994) Competing paradigms in qualitative research. In N. K. Denzin and Y. S. Lincoln (eds) *Handbook of Qualitative Research*. Beverly Hills, CA: Sage, pp. 105–17.
Guba, E. G. and Lincoln, Y. S. (2005) Paradigmatic controversies, contradictions, and emerging confluences. In N. K. Denzin and Y. S. Lincoln (eds) *The Sage Handbook of Qualitative Research* (third edition). Thousand Oaks, CA: Sage, pp. 191–215.
Guilford, J. P. and Fruchter, B. (1973) *Fundamental Statistics in Psychology and Education*. New York: McGraw-Hill.
Gwartney, P. A. (2007) *The Telephone Interviewer's Handbook*. San Francisco, CA: Jossey-Bass.
Habermas, J. (1972) *Knowledge and Human Interests* (trans. J. Shapiro). London: Heinemann.
Habermas, J. (1974) *Theory and Practice* (trans. J. Viertel). London: Heinemann.
Habermas, J. (1976) *Legitimation Crisis* (trans. T. McCarthy). London: Heinemann.
Habermas, J. (1979) *Communication and the Evolution of Society* (trans. T. McCarthy). London: Heinemann.
Habermas, J. (1982) A reply to my critics. In J. Thompson and D. Held (eds) *Habermas*: *Critical Debates*. London: Macmillan, pp. 219–83.
Habermas, J. (1984) *The Theory of Communicative Action, Vol. 1: Reason and the Rationalization of Society* (trans. T. McCarthy). Boston: Beacon Press.
Habermas, J. (1987a) *The Philosophical Discourse of Modernity* (trans. F. Lawrence). Cambridge, MA: Massachusetts Institute of Technology.
Habermas, J. (1987b) *The Theory of Communicative Action, Vol. 2: Lifeworld and System* (trans. T. McCarthy). Boston, MA: Beacon.
Habermas, J. (1988) *On the Logic of the Social Sciences* (trans. S. Nicholsen and J. Stark). Oxford: Polity Press in association with Basil Blackwell.
Habermas, J. (1990) *Moral Consciousness and Communicative Action* (trans. C. Lenhardt and S. Nicholsen). Cambridge: Polity Press in association with Basil Blackwell.
Hadfield, M. (2012) Becoming critical again: reconnecting critical social theory with the practice of action research. *Educational Action Research*, 20 (4), pp. 571–85.
Hadfield, M. and Haw, K. (2012) VideoL: modalities and methodologies. *International Journal of Research and Method in Education*, 35 (3), pp. 311–24.
Hage, J. and Meeker, B. F. (1988) *Social Causality*. London: Unwin Hyman Ltd.
Haggerty, K. (2004) Ethics creep: governing social science research in the name of ethics. *Qualitative Sociology*, 27 (4), pp. 391–414.
Haig, B. D. (1997) Feminist research methodology. In J. P. Keeves (ed.) *Educational Research, Methodology, and Measurement: An International Handbook* (second edition). Oxford: Elsevier Science, pp. 180–5.
Haig, B. D. (1999) Feminist research methodology. In J. P. Keeves and G. Lakomski (eds) *Issues in Educational Research*. Oxford: Elsevier Science Ltd, pp. 222–31.
Haladyna, T. M. (1997) *Writing Test Items to Evaluate Higher Order Thinking*. Needham Heights, MA: Allyn & Bacon.
Haladyna, T., Nolen, S. and Haas, N. (1991) Raising standardised achievement test scores and the origins of test score pollution. *Educational Researcher*, 20 (5), pp. 2–7.
Hall, B. (2005) Breaking the monopoly of knowledge: research methods, participation and development. In R. Tandon (ed.) *Participatory Research: Revisiting the Roots*. New Delhi: Mosaic Books, pp. 9–21.
Hall, S. (1996) Reflexivity in emancipatory action research: illustrating the researcher's constitutiveness. In O. Zuber-Skerritt (ed.) *New Directions in Action Research*. London: Falmer, pp. 26–48.
Hallett, R. E. and Barber, K. (2014) Ethnographic research in a cyber era. *Journal of Contemporary Ethnography*, 43 (3), pp. 306–30.
Halperin, D. M. (1997) *Saint Foucault: Towards a Gay Hagiography*. New York: Oxford University Press.
Hambleton, R. K. (1993) Principles and selected application of item response theory. In R. Linn (ed.) *Educational Measurement* (third edition). Phoenix, AZ: American Council on Education and the Oryx Press, pp. 147–200.
Hambleton, R. K. (2012) Measurement and validity. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 241–7.
Hamilton, L. and Corbett-Whittier, C. (2013) *Using Case Study in Education Research*. London: Sage.
Hamilton, M. L., Smith, L. and Worthington, K. (2008) Fitting the methodology with the research: an exploration of narrative, self-study and auto-ethnography. *Studying Teacher Education*, 4 (1), pp. 17–28.
Hammersley, M. (1992a) Deconstructing the qualitative–quantitative divide. In J. Brannen (ed.) *Mixing Methods: Qualitative and Quantitative Research*. Aldershot, UK: Avebury, pp. 39–57.
Hammersley, M. (1992b) *What's Wrong with Ethnography?* London: Routledge.
Hammersley, M. (2000) *Taking Sides in Social Research: Essays on Bias and Partisanship*. London: Routledge.
Hammersley, M. (2001) On 'systematic' reviews of research literatures: a 'narrative' response to Evans and Benefield. *British Educational Research Journal*, 27 (5), pp. 543–54.
Hammersley, M. (2003) Systematic or unsystematic? Is that the question? Some reflections on the science, art, and politics of reviewing research. Paper presented at the Department of Epidemiology and Public Health, University of Leicester, UK, February.
Hammersley, M. (2004) Some questions about evidence-based practice in education. In G. Thomas and R. Pring (eds) *Evidence-Based Practice in Education*. Buckingham, UK: Open University Press, pp. 133–49.
Hammersley, M. (2005) Is the evidence-based practice movement doing more good than harm? Reflections on Iain Chalmers' case for research-based policymaking and practice. *Evidence and Policy*, 1 (1), pp. 1–16.
Hammersley, M. (2006) Ethnography: problems and prospects. *Ethnography and Education*, 1 (1), pp. 3–14.
Hammersley, M. (2007) The issue of quality in qualitative research. *International Journal of Research and Method in Education*, 30 (3), pp. 287–305.

Hammersley, M. (2008) Paradigm war revived? On the diagnosis of resistance to randomized controlled trials and systematic review in education. *International Journal of Research and Method in Education*, 31 (1), pp. 3–10.

Hammersley, M. (2009) Against the ethicists: on the evils of ethical regulation. *International Journal of Social Research Methodology*, 12 (3), pp. 211–25.

Hammersley, M. (2011) *Methodology: Who Needs It?* London: Sage.

Hammersley, M. (2012) Troubling theory in case study research. *Higher Education Research and Development*, 31 (3), pp. 393–405.

Hammersley, M. (2013) *What Is Qualitative Research?* London: Bloomsbury Academic.

Hammersley, M. (2014) *The Limits of Social Science: Causal Explanation and Value Relevance*. London: Sage.

Hammersley, M. (2015a) Against 'gold standards' in research: on the problem of assessment criteria. Paper given at 'Was heißt hier eigentlich "Evidenz"?', Frühjahrstagung 2015 des AK Methoden in der Evaluation Gesellschaft für Evaluation (DeGEval), Fakultät für Sozialwissenschaften, Hochschule für Technik und Wirtschaft des Saarlandes, Saarbrücken, May 2015. Available from: www.degeval.de/fileadmin/users/Arbeitskreise/AK_Methoden/Hammersley_Saarbrucken.pdf [Accessed 10 April 2016].

Hammersley, M. (2015b) On ethical principles for social research. *International Journal of Social Research Methodology*, 18 (4), pp. 433–49.

Hammersley, M. and Atkinson, P. (1983) *Ethnography: Principles in Practice*. London: Routledge.

Hammersley, M. and Traianou, A. (2012) *Ethics in Qualitative Research: Controversies and Contexts*. London: Sage.

Hampden-Thompson, G., Lubben, F. and Bennett, J. (2011) Post-16 physics and chemistry uptake: combining large-scale secondary analysis with in-depth qualitative methods. *International Journal of Research and Method in Education*, 34 (3), pp. 289–307.

Hampden-Turner, C. (1970) *Radical Man*. Cambridge, MA: Schenkman.

Haney, C. and Zimbardo, P. G. (1998) The past and future of U.S. prison policy: twenty-five years after the Stanford Prison Experiment. *American Psychologist*, 53 (7), pp. 709–27.

Hanna, G. S. (1993) *Better Teaching through Better Measurement*. Fort Worth, TX: Harcourt Brace Jovanovich Inc.

Hanna, P. (2012) Using internet technologies (such as Skype) as a research medium: a research note. *Qualitative Research*, 12 (2), pp. 239–42.

Hannes, K. and Lockwood, C. (2011a) Pragmatism as the philosophical foundation for the Joanna Briggs meta-aggregative approach to qualitative evidence synthesis. *Journal of Advanced Nursing*, 67 (7), pp. 1632–42.

Hannes, K. and Lockwood, C. (2011b) *Synthesizing Qualitative Research: Choosing the Right Approach* (second edition). Hoboken, NJ: John Wiley & Sons.

Hannes, K., Lockwood, C. and Pearson, A. (2010) A comparative analysis of three online appraisal instruments: ability to assess validity in qualitative research. *Qualitative Health Research*, 20 (12), pp. 1736–43.

Harding, D. J. and Seefeldt, K. S. (2013) Mixed methods and causal analysis. In S. L. Morgan (ed.) *Handbook of Causal Analysis for Social Research*. Dordrecht: Springer, pp. 91–110.

Hare, A. P. (1985) *Social Interactions as Drama: Applications from Conflict Resolution*. Beverly Hills, CA: Sage.

Haritos, A., Gindidis, A., Doan, C. and Bell, R. C. (2004) The effect of element role titles on construct structure and content. *Journal of Constructivist Psychology*, 17 (3), pp. 221–36.

Harlen, W. (ed.) (1994) *Enhancing Quality in Assessment*. London: Paul Chapman Publishing.

Harlow, A. (2010) Online surveys: possibilities, pitfalls and practicalities – the experience of the TELA evaluation. *Waikoto Journal of Education*, 15 (2), pp. 95–108.

Harper, D. (2000) Reimagining visual methods: Galileo to Neuromancer. In N. K. Denzin and Y. S. Lincoln (eds) *Handbook of Qualitative Research* (second edition). London: Sage, pp. 717–32.

Harper, D. (2002) Talking about pictures: a case for photo elicitation. *Visual Studies*, 17 (1), pp. 13–26.

Harré, R. (1972) *The Philosophies of Science*. Oxford: Oxford University Press.

Harré, R. (1976) The constructive role of models. In L. Collins (ed.) *The Use of Models in the Social Sciences*. London: Tavistock Publications, pp. 16–43.

Harré, R. and Secord, P. (1972) *The Explanation of Social Behaviour*. Oxford: Basil Blackwell.

Harris, N., Pearce, P. and Johnstone, S. (1992) *The Legal Context of Teaching*. London: Longman.

Harrison, R. and Stokes, H. (1992) *Diagnosing Organizational Culture*. San Francisco, CA: Jossey-Bass.

Hartley, J. and Betts, L. R. (2010) Four layouts and a finding: the effects of changes in the order of the verbal labels and numerical values on Likert-type scales. *International Journal of Social Research Methodology*, 13 (1), pp. 17–27.

Harvey, C. D. H. (1988) Telephone survey techniques. *Canadian Home Economics Journal*, 38 (1), pp. 30–5.

Hassey, N. (2015) Randomised control trials and their limitations for use within educational research. *researchED*. Available from: www.workingoutwhatworks.com/en-GB/Magazine/2015/1/RCTs_and_their_limitations [Accessed 4 June 2016].

Hatten, K., Forin, T. R. and Adams, R. (2013) A picture elicits a thousand meanings: photo elicitation as a method for investigating cross-disciplinary identity development. Paper #7360 presented at the 120th ASEE Annual Conference and Exposition. Washington, DC: American Society for Engineering Education.

Hattie, J. (2009) *Visible Learning: A Synthesis of over 800 Meta-analyses Relating to Achievement*. London: Routledge.

Hattie, J., Rogers, H. J. and Swaminathan, H. (2014) The role of meta-analysis in educational research. In A. D. Reid, E. P. Hart and M. A. Peters (eds) *A Companion to Research in Education*. New York: Springer, pp. 197–207.

Hawkins, K. A. (2015) The complexities of participatory action research and the problems of power, identify and influence. *Educational Action Research*, 23 (4), pp. 464–78.

Haynes, L., Service, O., Goldacre, B. and Torgerson, D. (2012) *Test, Learn, Adapt: Randomised Controlled Trials*. London: Cabinet Office Behavioural Insights Team.

Head, E. (2009) The ethics and implications of paying participants in qualitative research. *International Journal of Social Research Methodology*, 12 (4), pp. 334–44.

Healy, K. (2001) Participatory action research and social work. *International Social Work*, 4 (1), pp. 93–105.

Heath, C. and Hindmarsh, J. (2002) Analysing interaction: video, ethnography and situated conduct. In T. May (ed.) *Qualitative Research in Action*. London: Sage, pp. 99–120.

Heath, C., Hindmarsh, J. and Luff, P. (2010) *Video in Qualitative Research: Analysing Social Interaction in Everyday Life*. London: Sage.

Heath, D. (2009) *The Literature Review: A Few Tips on Conducting It*. Health Sciences Writing Center, University of Toronto. 2010 *Literature Reviews*. Available from: www.writing.utoronto.ca/advice/specific-types-of-writing/literature-review [Accessed 6 February 2009].

Heathcote, D. (1991) *Collected Writings on Education and Drama* (ed. L. Johnson and C. O'Neill). Evanston, IL: Northwestern University Press.

Heaton, J. (1998) Secondary analysis of qualitative data. *Social Research Update 22*. Guildford, UK: University of Surrey, pp. 1–6.

Heaton, J. (2008) Secondary analysis of qualitative data: an overview. *Historical Social Research*, 31 (3), pp. 33–45.

Heck, R. H., Thomas, S. K. and Tabaska, L. N. (2013) *Multilevel and Longitudinal Modeling with IBM SPSS* (second edition). London: Routledge.

Heckathorn, D. D. (1997) Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems*, 44 (2), pp. 174–99.

Heckathorn, D. D. (2002) Respondent-driven sampling II: deriving population estimates from chain-referral samples of hidden populations. *Social Problems*, 49 (1), pp. 11–34.

Heckmann, M. (2014) *OpenRepGrid: An R Package for the Analysis of Repertory Grids*. R Package Version 0.1.9. Available from: https://cran.r-project.org/package=OpenRepGrid [Accessed 4 October 2016].

Heckmann, M. (2016) *The OpenRepGrid project: A Collection of Tools for the Analysis of Repertory Grid Data.* Available from: http://openrepgrid.org [Accessed 4 October 2016].

Heckmann, M. and Bell, R. C. (2015) Using linear mixed models with repertory grid data. In D. Winter and N. Reed (eds) *The Wiley-Blackwell Handbook of Personal Construct Psychology*. Chichester, UK: John Wiley, pp. 99–112.

Hedges, L. V. (1981) Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6 (2), pp. 107–28.

Hedges, L. V. and Olkin, I. (1985) *Statistical Methods for Meta-analysis*. Orlando, FL: Academic Press.

Heeran-Flynn, L. (2010) 'Breaking the code': an investigation of the use of drama in education in the development of pupils' oral language skills. Unpublished PhD thesis, Trinity College Dublin.

Heerwegh, D., Vanhove, T., Matthijs, K. and Loosveldt, G. (2005) The effect of personalization on response rates and data quality in web surveys. *International Journal of Social Research Methodology*, 8 (2), pp. 85–99.

Heikkinen, H. L. T., Huttunen, R., Syrjäla, L. and Pesonen, J. (2012) Action research and narrative inquiry: five principles for validation revisited. *Educational Action Research*, 20 (1), pp. 5–21.

Held, D. (1980) *Introduction to Critical Theory*. Los Angeles: University of California Press.

Hellevik, O. (1988) *Introduction to Causal Analysis*. Oslo: Norwegian University Press.

Helskog, G. H. (2014) Justifying action research. *Educational Action Research*, 22 (1), pp. 4–20.

Henderson, M., Johnson, N. F. and Auld, G. (2013) Silences of ethical practice: dilemmas for researchers using social media. *Educational Research and Evaluation*, 19 (6), pp. 546–60.

Hendry, P. M. (2007) The future of narrative. *Qualitative Inquiry*, 13 (4), pp. 487–98.

Hernandez, C. A. and Andrews, T. (2012) Commentary on 'Constructing new theory for identifying students with emotional disturbance'. *The Grounded Theory Review*, 11 (2), pp. 59–63.

Heron, J. and Reason, P. (1997) A participatory inquiry paradigm. *Qualitative Inquiry*, 3 (3), pp. 274–94.

Hertel, J. P. and Millis, B. J. (2002) *Using Simulations to Promote Learning in Higher Education*. Sterling, VA: Stylus Publishing.

Hesse, M. (1982) Science and objectivity. In J. Thompson and D. Held (eds) *Habermas: Critical Debates*. London: Macmillan, pp. 98–115.

Hesse-Biber, S. (2010) Qualitative approaches to mixed methods practice. *Qualitative Inquiry*, 16 (6), pp. 455–68.

Hesse-Biber, S. and Johnson, R. B. (2013) Coming at things differently: future directions of possible engagement with mixed methods research. *Journal of Mixed Methods Research*, 7 (2), pp. 103–9.

Hewson, C., Yule, P., Laurent, D. and Vogel, C. (2003) *Internet Research Methods*. London: Sage.

Heyvaert, M. (2013) Mixed methods research synthesis: definition, framework, and potential. *Quality and Quantity*, 47 (2), pp. 659–76.

Higgins, J. M. and McAllaster, C. (2004) If you want strategic change, don't forget to change your cultural artifacts. *Journal of Change Management*, 4 (1), pp. 63–73.

Hildenbrand, B. (2004) *Anselm Strauss*. In U. Flick, E. von Kardoff and I. Steinke (eds) *A Companion to Qualitative Research*. London: Sage, pp. 17–23.

Hill, H. C., Charalambous, C. Y. and Kraft, M. A. (2012) When rater reliability is not enough: teacher observation systems and a case for generalizability study. *Educational Researcher*, 41 (2), pp. 56–64.

Hillery, G. (1955) Definitions of community: areas of agreement. *Rural Sociology*, 20 (4), pp. 111–22.

Hilton, A. and Skrutkowski, M. (2002) Translating instruments into other languages: development and testing processes. *Cancer Nursing*, 25 (1), pp. 1–7.

Hilton, C. E. (2017) The importance of pretesting questionnaires: a field research example of cognitive pretesting the Exercise Referral Quality of Life Scale (ER-QLS). *International Journal of Social Research Methodology*, 20 (1), pp. 21–34.

Hinchcliffe, V. and Gavin, H. (2008) Internet mediated research: a critical reflection upon the practice of using instant messaging for higher educational research interviewing. *Psychology and Society*, 1 (1), pp. 91–104.

Hine, C. (2000) *Virtual Ethnography*. London: Sage.
Hine, C. (2004) *Virtual Ethnography Revisited*. Available from: www.restore.ac.uk/orm/background/exploringorms/rmf_hine_outline.pdf [Accessed 23 April 2016].
Hine, C. (2007) Multi-sited ethnography as a middle range methodology for contemporary STS. *Science, Technology and Human Values*, 32 (6), pp. 652–71.
Hine, C. (2015) *Ethnography for the Internet: Embedded, Embodied and Everyday*. London: Bloomsbury.
Hinkle, D. N. (1965) The change of personal constructs from the viewpoint of a theory of implications. Unpublished PhD thesis, Ohio State University.
Hinrichs, R. and Wankel, C. (2011) *Transforming Virtual World Learning: Cutting-Edge Technologies in Higher Education, Vol. 4*. London: Emerald.
Hitchcock, C. (2002) Probabilistic causation. In *Stanford Encyclopedia of Philosophy*. Available from: http://plato.stanford.edu/entried/causation-probabilistic [Accessed 31 December 2007].
Hitchcock, C. (ed.) (2004) *Contemporary Debates in Philosophy of Science*. Oxford: Blackwell Publishing.
Hitchcock, G. and Hughes, D. (1989) *Research and the Teacher*. London: Routledge.
Hitchcock, G. and Hughes, D. (1995) *Research and the Teacher* (second edition). London: Routledge.
Hobsbawm, E. (1964) *Labouring Men*. London: Weidenfield & Nicolson.
Hochschild, A. (2012) *The Managed Heart: Commercialization of Human Feeling* (revised third edition). Berkeley, CA: University of California Press.
Hochschild, J. L. (2009) Conducting intensive interviews and elite interviews. Workshop on Interdisciplinary Standards for Systematic Qualitative Research. Available from: http://scholar.harvard.edu/jlhochschild/publications/conducting-intensive-interviews-and-elite-interviews [Accessed 6 April 2016].
Hofstede, G. H. (1980) *Culture's Consequences: International Differences in Work-Related Values*. Beverly Hills, CA: Sage.
Hofstede, G. H. and Bond, M. H. (1984) Hofstede's cultural dimensions: an independent validation using Rokeach's Value Survey. *Journal of Cross-Cultural Psychology*, 15 (4), pp. 417–33.
Hoinville, G. and Jowell, R. (1978) *Survey Research Practice*. London: Heinemann.
Holbrook, D. (1977) *Education, Nihilism and Survival*. London: Darton, Longman & Todd.
Holland, J. H. and Miller, J. H. (1991) Artificial adaptive agents in economic theory. *American Economic Review*, 81 (2), pp. 356–71. Available from: http://ideas.repec.org/s/aea/aecrev17.html [Accessed 20 May 2010].
Holland, P. W. (1986) Statistics and causal inference. *Journal of the American Statistics Association*, 81, pp. 945–70.
Holly, P. (1984) *Action Research: A Cautionary Note*. Classroom Action Research Network, Bulletin No. 6. Cambridge: Cambridge Institute of Education.
Holly, P. and Whitehead, D. (1986) *Action Research in Schools: Getting It into Perspective*. Cambridge: Classroom Action Research Network.
Holmes, R. M. (1998) *Fieldwork with Children*. London: Sage.
Homer, M., Ryder, J. and Donnelly, J. (2011) The use of national data sets to baseline science education reform: exploring value-added approaches. *International Journal of Research and Method in Education*, 34 (3), pp. 309–25.
Hong, E., Mason, E., Peng, Y. and Lee, N. (2015) Effects of homework motivation and worry anxiety on homework achievement in mathematics and English. *Educational Research and Evaluation*, 21 (7–8), pp. 491–514.
Hopkins, D. (1985) *A Teacher's Guide to Classroom Research*. Milton Keynes, UK: Open University Press.
Hopkins, K. D., Hopkins, B. R. and Glass, G. V. (1996) *Basic Statistics for the Behavioral Sciences* (third edition). Boston, MA: Allyn & Bacon.
Horkheimer, M. (1972) *Critical Theory: Selected Essays* (trans. M. Connell). New York: Herder & Herder.
Hornsby-Smith, M. (1993) Gaining access. In N. Gilbert (ed.) *Researching Social Life*. London: Sage, pp. 52–67.
Horwich, P. (1993) Lewis's programme. In E. Sosa and M. Tooley (eds) *Causation*. Oxford: Oxford University Press, pp. 208–16.
Houssart, J. and Evens, H. (2011) Conducting task-based interviews with pairs of children: consensus, conflict, knowledge construction and turn taking. *International Journal of Research and Method in Education*, 34 (1), pp. 63–79.
Houtkook-Steenstra, H. and van den Bergh, H. (2000) Effects of introductions in large-scale telephone interviews. *Sociological Methods and Research*, 28 (3), pp. 281–300.
Howe, K. R. (1988) Against the quantitative–qualitative incompatibility thesis (or dogmas die hard). *Educational Researcher*, 17 (8), pp. 42–61.
Howe, K. R. and Moses, M. S. (1999) Ethics in educational research. *Review of Research in Education*, 24 (1), pp. 21–59.
Howitt, D. and Cramer, D. (2005) *Introduction to Research Methods in Psychology*. Harlow, UK: Pearson Education Ltd.
Howitt, D. and Cramer, D. (2014) *Introduction to Research Methods in Psychology* (fourth edition). Harlow, UK: Pearson.
Hoyle, E. (1986) *The Politics of School Management*. Sevenoaks, UK: Hodder & Stoughton.
Huberman, A. M. and Miles, M. B. (1998) Data management and analysis methods. In N. K. Denzin and Y. S. Lincoln (eds) *Collecting and Interpreting Qualitative Materials*. London: Sage, pp. 179–211.
Hudson, J. M. and Bruckman, A. (2005) Using empirical data to reason about internet research ethics. In H. Gellersen, K. Schmidt, M. Beaudouin-Lafon and W. Mackay (eds) *ECSCW 2005: Proceedings of the 9th European Conference on Computer Supported Cooperative Work*. Rotterdam, Netherlands: Springer, pp. 289–306.
Hudson, P. and Miller, C. (1997) The treasure hunt: strategies for obtaining maximum response to a postal survey. *Evaluation and Research in Education*, 11 (2), pp. 102–12.
Huff, A. S. (2009) *Designing Research for Publication*. Thousand Oaks, CA: Sage.
Hughes, J. A. (1976) *Sociological Analysis: Methods of Discovery*. Sunbury-on-Thames, UK: Nelson and Sons.
Hult, M. and Lennung, S. (1980) Towards a definition of action-research: a note and bibliography. *Journal of Management Studies*, 17 (2), pp. 241–50.

Hume, D. (1955) *An Inquiry Concerning Human Understanding*. New York: Liberal Arts Press Inc.

Hume, D. (2000) *A Treatise of Human Nature* (ed. D. F. Norton and M. J. Norton). Oxford: Oxford University Press.

Humphreys, L. (1970) *Tearoom Trade: A Study of Homosexual Encounters in Public Places*. London: Gerald Duckworth.

Humphreys, L. (1975) *Tearoom Trade: Impersonal Sex in Public Places* (enlarged edition). New York: Aldine.

Hunsinger, J. and Krotoski, A. (eds) (2012) *Learning and Research in Virtual Worlds*. London: Routledge.

Hunter, J. E., Schmidt, F. L. and Jackson, G. B. (1982) *Meta-analysis: Cumulating Research Findings across Studies*. Beverly Hills: Sage.

Hurworth, R. (2003) Photo-interviewing for research. *Social Research Update*, 40, pp. 1–4.

Hurworth, R. (2012) Techniques to assist with interviewing. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 177–85.

Hustler, D., Edwards, A. and Stronach, I. (1998) Editorial. *British Educational Research Journal*, 24 (5), pp. 499–501.

Hutchinson, B. and Whitehouse, P. (1986) Action research, professional competence and school organization. *British Educational Research Journal*, 12 (1), pp. 85–94.

Hutchison, A. J., Johnston, L. H. and Breckon, J. D. (2010) Using QSR-NVivo to facilitate the development of a grounded theory project: an account of a worked example. *International Journal of Social Research Methodology*, 13 (4), pp. 283–302.

Hutchison, D. and Styles, B. (2010) *A Guide to Running Randomised Controlled Trials for Educational Researchers*. Slough, UK: National Foundation for Educational Research.

Hycner, R. H. (1985) Some guidelines for the phenomenological analysis of interview data. *Human Studies*, 8 (3), pp. 279–303.

Hydén, L. C. and Bülow, P. H. (2003) Who's talking: drawing conclusions from focus groups – some methodological considerations. *International Journal of Social Research Methodology*, 6 (5), pp. 305–21.

Ijsselsteijn, W. A., de Ridder, H., Freeman, J. and Avons, S. E. (2000) Presence: concept, determinants and measurement. *Proceedings of the SPIE*, 3959, pp. 520–9.

INCITE (2010) *Participatory Action Research*. Redmond, WA: INCITE. Available from: www.incite-national.org/media/docs/5614_toolkitrev-par.pdf [Accessed 24 April 2010].

International Council on Human Rights Policy (2011) *Navigating the Dataverse: Privacy, Technology, Human Rights*. Geneva, Switzerland: International Council on Human Rights.

Ions, E. (1977) *Against Behaviouralism: A Critique of Behavioural Science*. Oxford: Basil Blackwell.

Ivankova, N. V. (2013) Implementing quality criteria in designing and conducting a sequential QUAN®QUAL mixed methods study of student engagement with learning applied research methods online. *Journal of Mixed Methods Research*, 8 (1), pp. 25–51.

Ivankova, N. V., Creswell, J. W. and Stick, S. (2006) Using mixed methods sequential explanatory design: from theory to practice. *Field Methods*, 18 (1), pp. 3–20.

Izard, J. (2005) Overview of test construction. *Module 6: Quantitative Research Methods in Educational Planning*. Paris: International Institute for Educational Planning, UNESCO.

Jackson, G. B. (1980) Methods for integrative review. *Review of Educational Research*, 50 (3), pp. 438–60.

Jackson, P. W. (1968) *Life in Classrooms*. New York: Holt, Rinehart & Winston.

James, M. (1993) Evaluation for policy: rationality and political reality – the paradigm case of PRAISE? In R. G. Burgess (ed.) *Educational Research and Evaluation for Policy and Practice*. London: Falmer, pp. 119–38.

James, N. (2007) The use of email interviewing as a qualitative method of inquiry in educational research. *British Educational Research Journal*, 33 (6), pp. 963–76.

James, N. (2015) You've got mail …! Using email interviews to gather academics' narratives of their working lives. *International Journal of Research and Method in Education*. DOI: 10.1080/1743727X.2015.1056136.

James, N. (2016) Using email interviews in qualitative educational research: creating space to think and time to talk. *International Journal of Qualitative Studies in Education*, 29 (2), pp. 150–63.

James, N. and Busher, H. (2006) Credibility, authenticity and voice: dilemmas in online interviewing. *Qualitative Research*, 6 (3), pp. 403–20.

James, N. and Busher, H. (2007) Ethical issues in online educational research: protecting privacy, establishing authenticity in email interviewing. *International Journal of Research and Method in Education*, 30 (1), pp. 101–13.

James, N. and Busher, H. (2015) Editorial: ethical issues in online research. *Educational Research and Evaluation*, 21 (2), pp. 89–94.

Jameson, F. (1991) *Postmodernism, or the Cultural Logic of Late Capitalism*. London: Verso.

Jankowicz, D. (2003) *The Easy Guide to Repertory Grids.* Chichester, UK: John Wiley.

Jansen, A. (2015) Positioning and subjectivism in research interviews: why bother talking to a researcher? *International Journal of Social Research Methodology*, 18 (1), pp. 27–39.

Jarmon, L., Traphagan, T. W., Traphagan, J. W. and Eaton, L. J. (2009) Ageing, lifelong learning and the virtual world of Second Life. In C. Wankel and J. Kingsley (eds) *Higher Education in Virtual Worlds: Teaching and Learning in Second Life*. Bingley, UK: Emerald, pp. 221–42.

Jarzemsky, P., McCarthy, J. and Ellis, N. (2010) Incorporating quality and safety education for nurses competencies in simulation scenario design. *Nurse Educator*, 35 (2), pp. 90–2.

Jayaratne, T. E. (1993) The value of quantitative methodology for feminist research. In M. Hammersley (ed.) *Social Research: Philosophy, Politics and Practice*. London: Sage in association with the Open University Press, pp. 109–23.

Jayaratne, T. E. and Stewart, A. (1991) Quantitative and qualitative methods in the social sciences: current feminist issues and practical strategies. In M. Fonow and J. Cook

(eds) *Beyond Methodology: Feminist Scholarship as Lived Research*. Bloomington, IN: Indiana University Press, pp. 133–53.

Jefferson, R. N. (2014) Action research: theory and application. *New Review of Academic Librarianship*, 20 (2), pp. 91–116.

Jennings, P. A., Frank, J. L., Snowberg, K. E., Coccia, M. A. and Greenberg, M. T. (2013) Improving learning environments by Cultivating Awareness and Resilience in Education (CARE): results of a randomized controlled trial. *School Psychology Quarterly*, 28 (4), pp. 374–90.

Jensen, L. A. and Allen, M. N. (1994) A synthesis of qualitative research on wellness-illness. *Qualitative Health Research*, 4 (4), pp. 349–69.

Jensen, L. A. and Allen, M. N. (1996) Meta-synthesis of qualitative findings. *Qualitative Health Research*, 6 (4), pp. 553–60.

Jewitt, C. (2012) *An Introduction to Using Video for Research*. National Centre for Research Methods Working Paper 03/12. London: National Centre for Research Methods.

Johansson, J., Skeff, K. M. and Stratos, G. A. (2012) A randomised control study of role play in a faculty development programme. *Medical Teacher*, 34 (2), pp. 123–8.

Johns, M. D., Chen, S. S. and Hall, G. J. (eds) (2004) *Online Social Research: Methods, Issues and Ethics*. New York: Peter Lang.

Johnson, A. and Sackett, R. (1998) Direct systematic observation of behavior. In H. R. Bernard (ed.) *Handbook of Methods in Cultural Anthropology*. Walnut Creek, CA: Altamira Press, pp. 301–31.

Johnson, B. and Christensen, L. (2010) *Educational Research: Quantitative, Qualitative, and Mixed Approaches*. Beverly Hills, CA: Sage.

Johnson, M. (2008) Assessing at the borderline: judging a vocationally related portfolio holistically. *Issues in Educational Research*, 18 (1), pp. 26–43.

Johnson, M. and Black, B. (2012) What's going on? Analysing visual data to understand context-based decision-making processes. *International Journal of Research and Method in Education*, 35 (3), pp. 243–50.

Johnson, R. B. and Onwuegbuzie, A. J. (2004) Mixed methods research: a research paradigm whose time has come. *Educational Researcher*, 33 (7), pp. 14–26.

Johnson, R. B., Onwuegbuzie, A. J. and Turner, L. (2007) Towards a definition of mixed methods research. *Journal of Mixed Methods Research*, 1 (2), pp. 112–33.

Johnston, M. P. (2014) Secondary data analysis: a method of which the time has come. *Qualitative and Quantitative Methods in Libraries*, 3 (3), pp. 619–26.

Joinson, A. N. and Reips, U.-D. (2007) Personal salutation, power of sender and response rates to Web-based surveys. *Computers in Human Behavior*, 23 (3), pp. 1372–83.

Joinson, A. N., McKenna, K., Postmes, T. and Reips, U.-D. (eds) (2009) *The Oxford Handbook of Internet Psychology*. Oxford: Oxford University Press.

Jones, C. (2011) Ethical issues in online research. *British Educational Research Association Online Resource*. Available from: www.bera.ac.uk/wp-content/uploads/2014/03/Ethical-issues-in-online-research.pdf?noredirect=1[Accessed 4 April 2016].

Jones, D. (1998) A biographical approach to the history of education: nineteenth century nonconformist lives and educational expansion. In M. Erben (ed.) *Biography and Education*. London: Falmer Press, pp. 130–48.

Jones, G. S. (1971) *Outcast London*. Oxford: Clarendon Press.

Jones, M. and Stanley, G. (2010) Collaborative action research: a democratic undertaking or a web of collusion and compliance. *International Journal of Research and Method in Education*, 33 (2), pp. 151–63.

Jones, N. (1999) The changing management agenda for primary heads. *International Journal of Public Sector Management*, 12 (4), pp. 324–37.

Jones, S. (1987) The analysis of depth interviews. In R. Murphy and H. Torrance (eds) *Evaluating Education: Issues and Methods*. London: Paul Chapman Publishing, pp. 263–77.

Jones, S. (ed.) (1999) *Doing Internet Research: Critical Issues and Methods for Examining the Net*. Thousand Oaks, CA: Sage.

Jordanova, L. (2000) *History in Practice*. London: Longman.

Joy, G. T. (2003) A brief description of culturally valid knowledge. Personal communication. Sophia Junior College, Hakone-machi, Kanagawa-ken, Japan.

Joyner, B. and Young, L. (2006) Teaching medical students using role-play: twelve tips for successful role plays. *Medical Teacher*, 28 (3), pp. 225–9.

Junco, R. (2012) Too much Face and not enough Books: the relationship between multiple indices of Facebook use and academic performance. *Computers in Human Behavior*, 28 (1), pp. 187–98.

Kahneman, D. (2012) *Thinking Fast and Slow*. London: Penguin.

Kaplan, A. (1964) *The Conduct of Inquiry*. San Francisco, CA: Chandler.

Kaplan, A. and Haenlein, M. (2010) Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53 (1), pp. 59–68.

Kapoor, D. and Jordan, S. (2009) *Education, Participatory Action Research, and Social Change*. New York: Palgrave Macmillan.

Karlsson, J. (2012) Visual methodologies. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 94–101.

Kauffman, S. A. (1993) *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford: Oxford University Press.

Kauffman, S. A. (1995) *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*. Harmondsworth: Penguin.

Kaufmann, J. J. (2011) Heteronarrative analysis: examining online photographic narratives. *International Journal of Qualitative Studies in Education*, 24 (1), pp. 7–26.

Kavanagh, K., Moro, T., Savage, T. and Mehendale, R. (2006) Enacting a theory of caring to recruit and retain vulnerable participants for sensitive research. *Research in Nursing and Health*, 29 (3), pp. 244–52.

Kawulich, B. B. (2005) Participant observation as a data collection method. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 6 (2), article 43. Available from: http://nbnresolving.de/urn:nbn:de:0114-fqs0502430 [Accessed 6 April 2010].

Kazdin, A. E. (1982) *Single-Case Research Designs*. New York: Oxford University Press.

Kazmer, M. M. and Xie, B. (2008) Qualitative interviewing in internet studies. *Information, Communication and Society*, 11 (2), pp. 257–78.

Kean, H. (1990) *Socialists and Feminists Challenging the State? The Socialist and Feminist Educational Experience 1900–1930*. London: Falmer Press.

Keane, E. (2015) Considering the practical implications of constructivist grounded theory in a study of widening participation in Irish higher education. *International Journal of Social Research Methodology*, 18 (4), pp. 415–31.

Keat, R. (1981) *The Politics of Social Theory*. Oxford: Basil Blackwell.

Kee, K. F. and Browning, L. D. (2013) Recruiting for and conducting qualitative telephone interviews to study dispersed groups, virtual organizations, and distributed communities. Paper presented at the National Communication Association annual conference. Washington, DC. Available from: www1.chapman.edu/~kee/PDF/C17.pdf [Accessed 6 April 2016].

Keet, H. M., Van Den Oord, E. J. C. G., Verhulst, F. C. and Boomsman, D. L. (1997) Behavioral and emotional problems in young preschoolers: cross-cultural testing of the validity of the Child Behavior Checklist/2–3. *Journal of Abnormal Child Psychology*, 25 (3), pp. 183–96.

Keeves, J. P. (1997a) Longitudinal research methods. In J. P. Keeves (ed.) *Educational Research, Methodology and Measurement: An International Handbook* (second edition). Oxford: Elsevier Science Ltd, pp. 138–49.

Keeves, J. P. (ed.) (1997b) *Educational Research, Methodology and Measurement: An International Handbook* (second edition). Oxford: Elsevier Science Ltd.

Keeves, J. P. and Sellin, N. (1997) Multilevel analysis. In J. P. Keeves (ed.) *Educational Research, Methodology and Measurement: An International Handbook* (second edition). Oxford: Elsevier Science Ltd, pp. 394–403.

Kelle, U. (ed.) (1995) *Computer-Aided Qualitative Data Analysis*. London: Sage.

Kelle, U. (1997) Theory building in qualitative research and computer programmes for the management of textual data. *Sociological Research Online*, 2 (2). Available from: www.socresonline.org.uk/2/2/1.html [Accessed 1 May 2010].

Kelle, U. (2000) Computer assisted analysis: coding and indexing. In M. Bauer and G. Gaskell (eds) *Qualitative Researching with Text, Image, and Sound*. London: Sage, pp. 282–98.

Kelle, U. (2004) Computer-assisted analysis of qualitative data. In U. Flick, E. von Kardoff and I. Steinke (eds) *A Companion to Qualitative Research* (trans. B. Jenner). London: Sage, pp. 276–93.

Kelle, U. and Laurie, H. (1995) Computer use in qualitative research and issues of validity. In U. Kelle (ed.) *Computer-Aided Qualitative Data Analysis*. London: Sage, pp. 19–28.

Kelly, A. (1978) Feminism and research. *Women's Studies International Quarterly*, 1 (3), pp. 225–32.

Kelly, A. (1985) Action research: what is it and what can it do? In R. G. Burgess (ed.) *Issues in Educational Research: Qualitative Methods*. Lewes, UK: Falmer, pp. 129–51.

Kelly, A. (1989) Education or indoctrination? The ethics of school-based action research. In R. G. Burgess (ed.) *The Ethics of Educational Research*. Lewes, UK: Falmer, pp. 100–13.

Kelly, B. (2007) Methodological issues for qualitative research with learning disabled children. *International Journal of Social Research Methodology*, 10 (1), pp. 21–35.

Kelly, G. A. (1955) *The Psychology of Personal Constructs*. New York: Norton.

Kelly, G. A. (1969) *Clinical Psychology and Personality: The Selected Papers of George Kelly* (ed. B. A. Maher). New York: John Wiley.

Kelly, S. and Allison, M. A. (1999) *The Complexity Advantage: How the Science of Complexity Can Help Your Business Achieve Peak Performance*. New York: McGraw-Hill.

Kelman, H. C. (1967) Human use of human subjects. *Psychological Bulletin*, 67 (1), pp. 1–11.

Kemmis, S. (1982) Seven principles for programme evaluation in curriculum development and innovation. *Journal of Curriculum Studies*, 14 (3), pp. 221–40.

Kemmis, S. (1997) Action research. In J. P. Keeves (ed.) *Educational Research, Methodology, and Measurement: An International Handbook* (second edition). Oxford: Elsevier Science Ltd, pp. 173–9.

Kemmis, S. (2006) Participatory action research and the public sphere. *Educational Action Research*, 14 (4), pp. 459–76.

Kemmis, S. (2009) Action research as a practice-based practice. *Educational Action Research*, 17 (3), pp. 463–74.

Kemmis, S. (2010) What is to be done? The place of action research. *Educational Action Research*, 18 (4), pp. 417–27.

Kemmis, S. and McTaggart, R. (eds) (1981) *The Action Research Planner* (first edition). Geelong, Victoria: Deakin University Press.

Kemmis, S. and McTaggart, R. (eds) (1988) *The Action Research Planner* (second edition). Geelong, Victoria: Deakin University Press.

Kemmis, S. and McTaggart, R. (1992) *The Action Research Planner* (third edition). Geelong, Victoria: Deakin University Press.

Kemmis, S. and McTaggart, R. (2005) Participatory action research: communicative action research and the public sphere. In N. Denzin and Y. Lincoln (eds) *The Sage Handbook of Qualitative Research* (third edition). Thousand Oaks, CA: Sage, pp. 651–79.

Kemmis, S., McTaggart, R. and Nixon, R. (2014) *The Action Research Planner: Doing Critical Participatory Action Research*. Dordrecht: Springer.

Kenett, R. S. (2006) On the planning and design of sample surveys. *Journal of Applied Statistics*, 33 (4), pp. 405–15.

Kennedy, M. M. (2007) Defining a literature. *Educational Researcher*, 36 (3), pp. 139–47.

Kennedy-Lewis, B. L., Murphy, A. S. and Groslad, T. J. (2016) Using narrative inquiry to understand persistently disciplined middle school students. *International Journal of Qualitative Studies in Education*, 29 (1), pp. 1–28.

Kerlinger, F. N. (1970) *Foundations of Behavioral Research*. New York: Holt, Rinehart & Winston.

Kerlinger, F. N. (1986) *Foundations of Behavioral Research* (third edition). New York: Holt, Rinehart & Winston.

Kerlinger, F. N. (1991) Science and behavioural research. In D. S. Anderson and B. J. Biddle (eds) *Knowledge for*

Policy: Improving Education through Research. London: Falmer, pp. 87–102.

Kettley, N. (2012) *Theory Building in Educational Research*. London: Continuum Books.

Kgaile, A. P. and Morrison, K. R. B. (2006) Measuring and targeting internal conditions for school effectiveness in the Free State of South Africa. *Educational Management, Administration and Leadership*, 34 (1), pp. 47–68.

Khot, S. (2005) Popular theatre. In R. Tandon (ed.) *Participatory Research: Revisiting the Roots*. New Delhi: Mosaic Books, pp. 313–20.

Kimmel, A. J. (1988) *Ethics and Values in Applied Social Research*. Beverly Hills, CA: Sage.

Kincaid, H. (2004) There are laws in the social sciences. In C. Hitchcock (ed.) *Contemporary Debates in Philosophy of Science*. Oxford: Blackwell Publishing, pp. 168–85.

Kincaid, H. (2009) Causation in the social sciences. In H. Beebee, C. Hitchcock and P. Mensies (eds) *The Oxford Handbook of Causation*. Oxford: Oxford University Press, pp. 726–43.

Kincheloe, J. and McLaren, P. (1994) Rethinking critical theory and qualitative research. In N. K. Denzin and Y. S. Lincoln (eds) *Handbook of Qualitative Research*. Beverly Hills, CA: Sage, pp. 105–17.

Kincheloe, J. L. (2003) *Teachers as Researchers: Qualitative Inquiry as a Path to Empowerment* (second edition). London: RoutledgeFalmer.

Kirk, J. and Miller, M. L. (1986) *Reliability and Validity in Qualitative Research*. Qualitative Research Methods Series, no. 1. Beverly Hills, CA: Sage.

Kirk, J. and Wall, C. (2011) *Work and Identity: Historical and Cultural Context*. Basingstoke, UK: Palgrave Macmillan.

Kirk, R. E. (1999) *Statistics: An Introduction*. London: Harcourt Brace.

Kirschner, P. A. (2015) Facebook as learning platform: argumentation superhighway or dead-end street? *Computers in Human Behavior*, 53 (1), pp. 621–5.

Kitwood, T. M. (1977) Values in adolescent life: towards a critical description. Unpublished PhD thesis, School of Education, University of Bradford.

Kleven, T. A. (1995) Reliabilitet som pedagogisk problem (trans.: Reliability as an educational problem). Mimeo for doctoral lecture, 17 February 1995. Oslo: Institute for Educational Research. Quoted in B. Brock-Utne (1996) Reliability and validity in qualitative research within education in Africa. *International Review of Education*, 42 (6), pp. 605–21.

Kline, P. (2000) *Handbook of Psychological Testing* (second edition). London: Routledge.

Kline, P. (2016) *A Handbook of Test Construction: Introduction to Psychometric Design*. New York: Routledge.

Kline, R. B. (2004) *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*. Washington, DC: American Psychological Association.

Kline, R. B. (2005a) *Principles and Practice of Structural Equation Modeling* (second edition). New York: Guilford Press.

Kline, T. J. B. (2005b) Classical test theory. In *Psychological Testing: A Practical Approach to Design and Evaluation*. Thousand Oaks, CA: Sage, pp. 91–106.

Kline, R. B. (2015) *Principles and Practice of Structural Equation modeling* (fourth edition). New York: Guilford Press.

Klockars, C. B. (1979) Dirty hands and deviant subjects. In C. B. Klockars and F. O'Connor (eds) *Deviance and Decency: The Ethics of Research with Human Subjects*. Beverly Hills, CA: Sage, pp. 261–82.

Knoblauch, H., Baer, A., Laurier, E., Petschke, S. and Schnettler, B. (2008) Visual analysis: new developments in the interpretative analysis of video and photography. *Forum Qualitative Forschung/Forum: Qualitative Research*, 9 (3), pp. 1–10.

Knoblauch, H., Schnettler, B., Raab, J. and Soeffner, H. G. (eds) (2006) *Video Analysis: Methodology and Methods*. Frankfurt: Peter Lang.

Knott, J. and Wildavsky, A. (1991) If dissemination is the solution, what is the problem? In D. S. Anderson and B. J. Biddle (eds) *Knowledge for Policy: Improving Education through Research*. London: Falmer, pp. 214–24.

Kogan, M. and Atkin, J. M. (1991) Special commissions and educational policy in the U.S.A. and U.K. In D. S. Anderson and B. J. Biddle (eds) *Knowledge for Policy: Improving Education through Research*. London: Falmer, pp. 245–58.

Kolakowski, L. (1978) *Main Currents of Marxism, Vol. 3: The Breakdown* (trans. P. S. Falla). Oxford: Clarendon Press.

Kolb, S. M. (2012) Grounded theory and the constant comparative method: valid research strategies for educators. *Journal of Emerging Trends in Educational Research and Policy Studies*, 3 (1), pp. 83–6.

Konecki, K. (2009) Teaching visual grounded theory. *Qualitative Sociology Review*, 5 (3), pp. 64–92. Available from: www.qualitativesociologyreview.org/ENG/archive_eng.php [Accessed 18 May 2010].

Kontopoulou, K. and Fox, A. (2015) Designing a consequentially-based study into the online support of preservice teachers in the UK. *Educational Research and Evaluation*, 21 (2), pp. 122–38.

Kozinets, R. V. (2002) The field behind the screen: using Netography for marketing research in online communities. *Journal of Marketing Research*, 39 (1), pp. 61–72.

Kozinets, R. V. (2010) *Netography: Doing Ethnographic Research Online*. London: Sage.

Krähenbühl, S. and Blades, M. (2006) The effect of interviewing techniques on young children's responses to questions. *Child Care, Health and Development*, 32 (3), pp. 321–31.

Kraus, R. (2008) You must participate: violating research ethical principles through role-play. *College Teaching*, 56 (3), pp. 131–6.

Krejcie, R. V. and Morgan, D. W. (1970) Determining sample size for research activities. *Educational and Psychological Measurement*, 30 (3), pp. 607–10.

Kress, T. M. (2011) Stepping out of the academic brew: using critical research to break down hierarchies of knowledge production. *International Journal of Qualitative Studies in Education*, 24 (3), pp. 267–83.

Krippendorp, K. (2004) *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage.

Krosnick, J. A. (1991) Response strategies for coping with the cognitive demands of attitude measurement in surveys. *Applied Cognitive Psychology*, 5 (3), pp. 213–36.

Krosnick, J. A. (1999) Survey research. *Annual Review of Psychology*, 50 (1), pp. 537–67.

Krosnick, J. A. and Alwin, D. F. (1987) An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, 51 (2), pp. 201–19.

Krosnick, J. A. and Presser, S. (2010) Question and questionnaire design. In P. V. Marsden and J. V. Wright (eds) *Handbook of Survey Research*. Bingley, UK: Emerald Group Publishing Ltd, pp. 263–313.

Krueger, J. (2001) Null hypothesis significance testing: on the survival of a flawed method. *American Psychologist*, 56 (1), pp. 16–26.

Krueger, R. A. (1988) *Focus Groups: A Practical Guide for Applied Research*. Beverly Hills, CA: Sage.

Kruger, R. A. and Casey, M. A. (2000) *Focus Groups: A Practical Guide for Applied Research* (third edition). Thousand Oaks, CA: Sage.

Kucuk, S. and Sahin, I. (2013) From the perspective of Community of Inquiry framework: an examination of Facebook uses by pre-service teachers as a learning environment. *The Turkish Online Journal of Educational Technology*, 12 (2), pp. 142–56.

Kuhn, L. (2007) Why utilize complexity principles in social inquiry? *World Futures*, 63 (3), pp. 156–75.

Kuhn, T. S. (1962) *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Kulavuz-Onal, D. and Vásquez, C. (2013) Reconceptualising fieldwork in a netnography of an online community of English language teachers. *Ethnography and Education*, 8 (2), pp. 224–38.

Kvale, S. (1996) *Interviews*. London: Sage.

Labaree, R. V. (2013) *Organizing Your Social Sciences Research Paper: Types of Research Designs*. USC Libraries Research Guides. Los Angeles, CA: University of Southern California. Available from: http://libguides.usc.edu/writingguide/researchdesigns [Accessed 9 October 2015].

Labov, W. (1969) The logic of non-standard English. In N. Keddie (ed.) *Tinker, Tailor …: the Myth of Cultural Deprivation*. Harmondsworth: Penguin, pp. 21–66.

Labov, W. (1972) The transformation of experience in narrative syntax. In W. Labov (ed.) *Language in the Inner City: Studies in the Black English Vernacular*. Philadelphia, PA: University of Pennsylvania Press, pp. 354–96.

Labus, A., Despotović-Zrakić, M., Radenković, B., Bogdanović, Z. and Radenkovic, M. (2015) Enhancing formal e-learning with edutainment on social networks. *Journal of Computer Assisted Learning*, 31 (6), pp. 592–605.

Lacey, C. (1970) *Hightown Grammar: The School as a Social System*. Manchester: Manchester University Press.

Lahlou, S. (2011) How can we capture the subject's perspective? An evidence-based approach for the social scientist. *Social Science Information*, 50 (4), pp. 607–55.

Laing, R. D. (1967) *The Politics of Experience and the Bird of Paradise*. Harmondsworth: Penguin.

Lakatos, I. (1970) Falsification and the methodology of scientific research programmes. In I. Lakatos and A. Musgrave (eds) *Criticism and the Growth of Knowledge*. London: Cambridge University Press, pp. 91–195.

Lakomski, G. (1999) Critical theory. In J. P. Keeves and G. Lakomski (eds) *Issues in Educational Research*. Oxford: Elsevier Science Ltd, pp. 174–83.

Landau, S. and Chis Ster, I. (2010) Cluster analysis: overview. In P. Peterson, E. Baker and B. McGaw (eds) *International Encyclopedia of Education* (third edition). Oxford: Elsevier Ltd, pp. 72–83.

Landfield, A. W. (1971) *Personal Construct Systems in Psychotherapy*. Lincoln, NE: University of Nebraska Press.

Lansing, J. B., Ginsburg, G. P. and Braaten, K. (1961) *An Investigation of Response Error*. Studies in Consumer Savings, no. 2. Urbana, IL: University of Illinois Bureau of Economic and Business Research.

Larsson, S. (2009) A pluralist view of generalization in qualitative research. *International Journal of Research and Method in Education*, 32 (1), pp. 25–38.

Lather, P. (1986a) Issues of validity in openly ideological research: between a rock and a soft place. *Interchange*, 17 (4), pp. 63–84.

Lather, P. (1986b) Research as praxis. *Harvard Educational Review*, 56 (3), pp. 257–77.

Lather, P. (1991) *Getting Smart: Feminist Research and Pedagogy within the Post Modern*. New York: Routledge.

Lather, P. (1993) Fertile obsession: validity after poststructuralism. *The Sociological Quarterly*, 34 (4), pp. 673–93.

Lather, P. (1996) Troubling clarity: the politics of accessible language. *Harvard Educational Review*, 66 (3), pp. 524–45.

Lather, P. (1999) To be of use: the work of reviewing. *Review of Educational Research*, 69 (1), pp. 2–7.

Lather, P. (2004) Critical inquiry in qualitative research: feminist and poststructural perspectives – science 'after truth'. In K. de Marrais and S. D. Lapan (eds) *Foundations for Research: Methods of Inquiry in Education and the Social Sciences*. Mahwah, NJ: Lawrence Erlbaum, pp. 203–15.

Laudan, L. (1990) *Science and Relativism*. Chicago, IL: University of Chicago Press.

Laurillard, D. (2012) *Teaching as a Design Science*. London: Routledge.

Lave, J. and Kvale, S. (1995) What is anthropological research? An interview with Jean Lave by Steiner Kvale. *International Journal of Qualitative Studies in Education*, 8 (3), pp. 219–28.

Law, J. (2004) *After Method: Mess in Social Science Research*. London: Routledge.

Layder, D. (1994) *Understanding Social Theory*. London: Sage.

Lazarsfeld, P. P. and Barton, A. (1951) Qualitative measurement in the social sciences: classification, typologies and indices. In D. P. Lerner and H. D. Lasswell (eds) *The Policy Sciences*. Stanford, CA: Stanford University Press, pp. 155–92.

Leander, K. M. and McKim, K. K. (2003) Tracing the everyday 'sightings' of adolescents on the internet: a strategic adaptation of ethnography across online and offline spaces. *Education, Communication and Information*, 3 (2), pp. 211–40.

Lechuga, V. M. (2012) Exploring culture from a distance: the utility of telephone interviews in qualitative research. *International Journal of Qualitative Studies in Education*, 25 (3), pp. 251–68.

LeCompte, M. and Preissle, J. (1993) *Ethnography and Qualitative Design in Educational Research* (second edition). London: Academic Press Ltd.

LeCompte, M., Millroy, W. L. and Preissle, J. (eds) (1992) *The Handbook of Qualitative Research in Education*. London: Academic Press Ltd.

Lee, D., Arthur, I. T. and Morrone, A. S. (2015) Using video surveillance footage to support validity of self-reported classroom data. *International Journal of Research and Method in Education*. DOI: 10.1080/1743727X.2015.1075496.

Lee, M. C. Y. (2016) Finding cultural harmony in interviewing: the wisdom of the middle way. *International Journal of Research and Method in Education*, 39 (1), pp. 38–57.

Lee, R. M. (1993) *Doing Research on Sensitive Topics*. London: Sage.

Lee, R. M. and Renzetti, C. M. (1993) The problems of researching sensitive topics: an overview and introduction. In C. Renzetti and R. M. Lee (eds) *Researching Sensitive Topics*. London: Sage, pp. 3–12.

Leech, N. L. and Onwuegbuzie, A. J. (2004) A proposed fourth measure of significance: the role of economic significance in educational research. *Evaluation and Research in Education*, 18 (3), pp. 179–98.

Leech, N. L. and Onwuegbuzie, A. J. (2009) A typology of mixed methods research designs. *Quantity and Quality*, 43 (2), pp. 265–75.

Leeson, C. (2014) Asking difficult questions: exploring research methods with children on painful issues. *International Journal of Research and Method in Education*, 37 (2), pp. 206–22.

Lehr, R. (1992) Sixteen S-squared over D-squared: a relation for crude sample size estimates. *Statistics in Medicine*, 11 (8), pp. 1099–102.

Lehrer, R. and Franke, M. L. (1992) Applying personal construct psychology to the study of teachers' knowledge of fractions. *Journal for Research in Mathematical Education*, 23 (3), pp. 223–41.

Lemke, J. (2001) *Toward Systemic Educational Change: Questions from a Complex Systems Perspective.* Cambridge, MA: New England Complex Systems Institute. Available from: www.necsi.org/events/cxedk16_3.html [Accessed 10 November 2001].

Lemke, J. (2007) Video epistemology in-and-outside the box: traversing attentional spaces. In R. Goldman, R. Pea, B. Barron and S. Derry (eds) *Video Research in the Learning Sciences*. Mahwah, NJ: Lawrence Erlbaum, pp. 39–52.

Lemke, J. L. (1989) *Using Language in the Classroom* (second edition). Oxford: Oxford University Press.

Lempert, L. B. (2007) Asking questions of the data: memo writing in the grounded theory tradition. In A. Bryant and K. Charmaz (eds) *The SAGE Handbook of Grounded Theory*. London: Sage, pp. 245–65.

Leong, F. T. L., Schmitt, N. and Lyons, B. J. (2012) Developing testable and important research questions. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf and K. J. Sher (eds) *APA Handbook of Research Methods in Psychology, Vol. 1: Foundations, Planning, Measures, and Psychometrics*. Washington, DC: American Psychological Association, pp. 119–32.

Leow, C. (2009) Conducting a rigorous quasi-experimental evaluation using a school district's existing student database. *International Journal of Research and Method in Education*, 32 (1), pp. 69–88.

Leshem, S. (2012) The group interview experience as a tool for admission to teacher education. *Education Research International*, article 876764. Available from: www.hindawi.com/journals/edri/2012/876764 [Accessed 4 April 2016].

Levin, H. M. (1991) Why isn't educational research more useful? In D. S. Anderson and B. J. Biddle (eds) *Knowledge for Policy: Improving Education through Research*. London: Falmer, pp. 70–8.

Levis-Rozalis, M. (2003) Evaluation and research: differences and similarities. *The Canadian Journal of Program Evaluation*, 18 (2), pp. 1–31.

Lewin, K. (1946) Action research and minority problems. *Journal of Social Issues*, 2 (4), pp. 34–46.

Lewin, K. (1948) *Resolving Social Conflicts*. New York: Harper.

Lewin, R. (1993) *Complexity: Life on the Edge of Chaos*. London: Phoenix.

Lewin, R. and Regine, B. (2000) *The Soul at Work: Listen, Respond, Let Go – Embracing Complexity Science for Business Success*. New York: Simon & Schuster.

Lewins, A. and Silver, C. (2009) *Choosing a CAQDAS Package*. Available from: http://eprints.ncrm.ac.uk/791/1/2009ChoosingaCAQDASPackage.pdf [Accessed 9 June 2016].

Lewis, A. (1992) Group child interviews as a research tool. *British Educational Research Journal*, 18 (4), pp. 413–21.

Lewis, D. (1974) *Assessment in Education*. London: University of London Press.

Lewis, J. (2006) Making order out of a contested disorder: the utilisation of online support groups in social science research. *Qualitative Researcher*, 3, pp. 4–7.

Lewis-Beck, M. S. (ed.) (1993) *Experimental Design and Methods: International Handbook of Quantitative Applications in the Social Sciences, Vol. 3*. London: Sage.

Liang, M. Y. (2012) Foreign lucidity in online role-playing games. *Computer Assisted Language Learning*, 25 (5), pp. 455–73.

Lie, R. and Witteveen, L. (2017) Visual informed consent: informed consent without forms. *International Journal of Social Research Methodology*, 20 (1), pp. 63–75.

Lieberman, E. S. (2005) Nested analysis as a mixed-method strategy for comparative research. *American Political Science Review*, 99 (3), pp. 435–52.

Lieberson, S. (1985) *Making It Count: The Improvement of Social Research and Theory*. Berkeley, Los Angeles and London: University of California Press.

Liebling, H. and Shah, S. (2001) Researching sensitive topics: investigations of the sexual abuse of women in Uganda and girls in Tanzania. *Law, Social Justice and Global Development.* Available from: www2.warwick.ac.uk/fac/soc/law/elj/lgd/2001_1/liebling [Accessed 20 May 2010].

Lietz, P. and Keeves, J. P. (1997) Cross-sectional research methods. In J. P. Keeves (ed.) *Educational Research, Methodology and Measurement: An International Handbook* (second edition). Oxford: Elsevier Science Ltd, pp. 138–49.

Liew, H. P. (2013) *Teach Yourself Cluster Analysis, Conjoint Analysis, and Econometrics*. Self-published volume from CreateSpace Independent Publishing Platform. Available from: www.amazon.com/Yourself-Analysis-Conjoint-Econometrics-Techniques/dp/1493530402 [Accessed 27 August 2016].

Light, R. J., Singer, J. and Willett, J. (1990) *By Design: Conducting Research on Higher Education*. Cambridge, MA: Harvard University Press.

Likert, R. (1932) *A Technique for the Measurement of Attitudes*. New York: Columbia University Press.

Limerick, B., Burgess-Limerick, T. and Grace, M. (1996) The politics of interviewing: power relations and accepting the gift. *International Journal of Qualitative Studies in Education*, 9 (4), pp. 449–60.

Lin, N. (1976) *Foundations of Social Research.* New York: McGraw-Hill.

Lincoln, Y. S. (1990) Toward a categorical imperative for qualitative research. In E. Eisner and A. Peshkin (eds) *Qualitative Inquiry in Educational Research: The Continuing Debate*. New York: Teachers College Press, pp. 277–95.

Lincoln, Y. S. and Guba, E. (1985) *Naturalistic Inquiry*. Beverly Hills, CA: Sage.

Lincoln, Y. S. and Guba, E. (1986) But is it rigorous? Trustworthiness and authenticity in naturalistic inquiry. In D. D. Williams (ed.) *Naturalistic Evaluation*. San Francisco, CA: Jossey-Bass, pp. 73–84.

Lindquist, E. F. (1940) *Statistical Analysis in Educational Research*. Boston, MA: Houghton Mifflin Company.

Linn, R. L. (ed.) (1993) *Educational Measurement* (third edition). Phoenix, AZ: American Council on Education and the Oryx Press.

Lipowski, E. E. (2008) Developing great research questions. *American Journal of Healthy Systems Pharmacists*, 65 (17), pp. 1667–70.

Lipsey, M. W. (1992) Juvenile delinquency treatment: a meta-analytic inquiry into the variability of effects. In T. D. Cook, H. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis and F. Mosteller (eds) *Meta-analysis for Explanation*. New York: Russell Sage Foundation, pp. 83–127.

Lipsey, M. W. and Wilson, D. B. (2001) *Practical Meta-analysis*. Thousand Oaks, CA: Sage.

Little, R. J. A. and Rubin, D. B. (1989) The analysis of social science data with missing values. *Sociological Methods and Research*, 6 (3), pp. 292–326.

Little, R. J. A. and Rubin, D. B. (2014) *Statistical Analysis with Missing Data* (second edition). Hoboken, NJ: John Wiley & Sons Inc.

Liu, H. J. C. (2002) Translation of instruments for cross-cultural research. *Journal of the Da-Yeh University*, 11 (2), pp. 79–88.

Livingston, G. (1999) Beyond watching over established ways: a review as recasting the literature, recasting the lived. *Review of Educational Research*, 69 (1), pp. 9–19.

Livingstone, D. and Bloomfield, P. R. (2010) Mixed-methods and mixed-worlds: engaging globally distributed user groups for extended evaluation and studies. In A. Peachey, J. Gillen, D. Livingstone and S. Smith-Robbins (eds) *Researching Learning in Virtual Worlds*. London: Springer, pp. 159–76.

Livingstone, I. (1999) Role-playing planning public inquiries. *Journal of Geography in Higher Education*, 23 (1), pp. 63–76.

Lobato, J. (2003) How design experiments can inform a rethinking of transfer and vice versa. *Educational Researcher*, 32 (1), pp. 17–20.

Locke, J. (1959) *An Essay Concerning Human Understanding, Vol. 1*. New York: Dover.

Locke, T., Alcorn, N. and O'Neill, J. (2013) Ethical issues in collaborative action research. *Educational Action Research*, 21 (1), pp. 107–23.

Lodico, M. G., Spaulding, D. T. and Voegtle, K. H. (2010) *Methods in Educational Research*. San Francisco, CA: Jossey-Bass.

Loehlin, J. (2004) *Latent Variable Models* (fourth edition). Mahwah, NJ: Lawrence Erlbaum Associates.

Loevinger, J. (1957) Objective tests as instruments of psychological theory. *Psychological Review*, 72, pp. 143–55.

Loewenthal, K. M. (2001) *An Introduction to Psychological Tests and Scales*. Hove: Psychology Press Ltd.

Lofland, J. (1970) Interactionist imagery and analytic interrupts. In T. Shibutani (ed.) *Human Nature and Collective Behaviour: Papers in Honour of Herbert Blumer*. Englewood Cliffs, NJ: Prentice-Hall, pp. 35–45.

Lofland, J. (1971) *Analysing Social Settings*. Belmont, CA: Wadsworth.

Long, H. Y. (2015) Validity in mixed methods research in education: the application of Habermas's critical theory. *International Journal of Research and Method in Education*. Available from: http://dx.doi.org/10.1080/1743727X.2015.1088518 [Accessed 8 June 2016].

Long-Sutenhall, T., Sque, M. and Addington-Hall, J. (2010) Secondary analysis of qualitative data: a valuable method for exploring sensitive issues with an elusive population. *Journal of Research in Nursing*, 16 (4), pp. 335–44.

Lonkila, M. (1995) Grounded theory as an emerging paradigm for computer-assisted qualitative data analysis. In U. Kelle (ed.) *Computer-Aided Qualitative Data Analysis*. London: Sage, pp. 41–51.

Lord, H. G. (1973) *Ex Post Facto Studies as a Research Method*. Special Report no. 7320. New York: Syracuse City School District. Available from: www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/39/5f/df.pdf [Accessed 10 April 2010].

Lowenstein, A. J. (2016) Role play. In M. J. Bradshaw and B. L. Hultquist (eds) *Innovative Teaching Strategies in Nursing and Related Health Professions*. Burlington, MA: Jones and Bartlett Publishers, Inc., pp. 211–28.

Lowenthal, D. (2015) *The Past Is a Foreign Country Revisited*. Cambridge: Cambridge University Press.

Lui, C. C. and Lee, J. H. (2005) Prompting conceptual understanding with computer-mediated peer discourse and knowledge acquisition techniques. *British Journal of Educational Technology*, 36 (5), pp. 821–37.

Lukenchuk, A. (2013) *Paradigms of Research for the 21st Century*. New York: Peter Lang.

Luttenberg, J., Meijer, P. and Oolbekkink-Marchand, H. (2016) Understanding the complexity of teacher reflection in action research. *Educational Action Research*. DOI: 10.1080/09650792.2015.1136230.

Lutz, C. A. and Collins, J. L. (1993) *Reading National Geographic*. Chicago, IL: University of Chicago Press.

McAteer, M. (2013) *Action Research in Education*. London: Sage.

McCandliss, B. D., Kalchman, M. and Bryant, P. (2003) Design experiments and laboratory approaches to learning:

steps towards collaborative exchange. *Educational Researcher*, 32 (1), pp. 14–16.

McCormick, R. and James, M. (1988) *Curriculum Evaluation in Schools* (second edition). London: Croom Helm.

McCosker, H., Barnard, A. and Gerber, R. (2001) Undertaking sensitive research: issues and strategies for meeting the safety needs of all participants. *Forum: Qualitative Social Research (Sozialforschung)*, 2 (1), pp. 1–10. Available from: www.qualitative-research.net/index.php/fqs/article/view/983 [Accessed 7 March 2016].

MacDonald, B. (1987) Evaluation and the control of education. In R. Murphy and H. Torrance (eds) *Evaluating Education: Issues and Methods*. London: Harper & Row, pp. 36–8.

McEwen, L., Stokes, A., Crowley, K. and Roberts, C. (2014) Using role-play for expert science communication with professional stakeholders in flood risk management. *Journal of Geography in Higher Education*, 38 (2), pp. 277–300.

McHugh, J. D. (1994) The Lords' will be done: interviewing the powerful in education. In G. Walford (ed.) *Researching the Powerful in Education*. London: UCL Press, pp. 51–66.

McKernan, J. (1991) *Curriculum Action Research*. London: Kogan Page.

Mackie, J. L. (1993) Causes and conditions. In E. Sosa and M. Tooley (eds) *Causation*. Oxford: Oxford University Press, pp. 33–55.

MacLure, M. (2005) 'Clarity bordering on stupidity': where's the quality in systematic review? *Journal of Education Policy*, 20 (4), pp. 393–416.

McNiff, J. (2010) *Action Research for Professional Development: Concise Advice for New and Experienced Action Researchers*. Poole, UK: September Books.

McNiff, J. and Whitehead, J. (2009) *Doing and Writing Action Research*. London: Sage.

McNiff, J., Lomax, P. and Whitehead, J. (1996) *You and Your Action Research Project*. London: Routledge, in association with Hyde Publications, Bournemouth.

Macpherson, I., Brooker, R. and Ainsworth, P. (2000) Case study in the contemporary world of research: using notions of purpose, place, process and product to develop some principles for practice. *International Journal of Research Methodology*, 3 (1), pp. 49–61.

McTaggart, R. (1989) *16 Tenets of Participatory Action Research*. Available from: www.caledonia.org.uk/par.htm [Accessed 24 April 2010].

McTaggart, R. (1996) Issues for participatory action researchers. In O. Zuber-Skerritt (ed.) *New Directions in Action Research*. London: Falmer, pp. 243–55.

Madden, N. A., Slavin, R. E., Logan, M. and Cheung, A. (2011) Effects of cooperative writing with embedded multimedia: a randomized experiment. *Effective Education*, 3 (1), pp. 1–9.

Madge, C. and O'Connor, H. (2005) Mothers in the making? Exploring notions of liminality in hybrid cyber/space. *Transactions in the Institute of British Geographers*, 30 (1), pp. 83–97.

Madge, J. (1965) *The Tools of Social Science*. London: Longman.

Madill, A. and Latchford, G. (2005) Identity change and the human dissection experience over the first year of medical training. *Social Science & Medicine*, 60 (7), pp. 1637–47.

Madison, D. S. (2005) *Critical Ethnography: Methods, Ethics and Performance*. London: Sage.

Madison, D. S. (2006) The dialogic performative in critical ethnography. *Text and Performance Quarterly*, 26 (4), pp. 320–4.

Magee, C., Rickards, G., Byars, L. A. and Artino, A. R., Jr (2013) Tracing the steps of survey design: a graduate medical education research example. *Journal of Graduate Medical Education*, 5 (10), pp. 1–5.

Mager, R. F. (1962) *Preparing Instructional Objectives*. Belmont, CA: Fearon Publishers.

Maguire, M. H. (2005) What if you talked to me? I could be interesting: ethical research considerations in engaging with bilingual/multicultural child participants in human inquiry. *Forum: Qualitative Sozialforschung/Forum: Qualitative Social Research*, 6 (1), pp. 1–24, article 4. Available from: www.qualitative-research.net/index.php/fqs/article/viewArticle/530/1148 [Accessed 28 March 2010].

Mahoney, J. (2000) Strategies of causal inference in small-n analysis. *Sociological Methods and Research*, 28 (4), pp. 387–424.

Maier, N. R. F., Solem, A. R. and Maier, A. A. (1957) *Supervisory and Executive Development: A Manual for Role-Playing*. Oxford: John Wiley.

Maines, B. and Robinson, G. (1997) *Crying for Help: The No Blame Approach to Bullying*. Bristol, UK: Lucky Duck Publishing.

Major, C. H. and Savin-Baden, M. (2010) *An Introduction to Qualitative Research Synthesis: Managing the Information Explosion in Social Science Research*. London: Routledge.

Malinowski, B. (1922) *Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagoes of Melanesian New Guinea*. New York: Dutton.

Maltese, A. V., Danish, J. A., Bouldin, R. M., Harsh, J. A. and Bryan, B. (2016) What are students doing during lecture? Evidence from new technologies to capture student activity. *International Journal of Research & Method in Education*, 39 (2), pp. 208–26.

Manca, S. and Ranieri, M. (2013) Is it a tool suitable for learning? A critical review of the literature on Facebook as a technology-enhanced learning environment. *Journal of Computer Assisted Learning*, 29 (6), pp. 487–504.

Manca, S. and Ranieri, M. (2016) Facebook and the others: potentials and obstacles of social media for teaching in higher education. *Computers and Education*, 95 (1), pp. 216–30.

Mannheim, K. (1936) *Ideology and Utopia*. London: Routledge & Kegan Paul.

Manton, K. (2001) *Socialism and Education in Britain 1883–1902*. London: Woburn Press.

Maratou, V., Chatzidaki, E. and Xenos, M. (2016) Enhance learning on software project management through a role-play game in a virtual world. *Interactive Learning Environments*, 24 (4), pp. 897–915.

Marion, R. (1999) *The Edge of Organization: Chaos and Complexity Theories of Formal Social Systems*. London: Sage.

Markham, A. (1998) *Life Online*. New York: Sage.

Markham, A. N. and Baym, N. K. (eds) (2008) *Internet Inquiry: Conversations about Method*. London: Sage.

Markus, H. R. and Kitayama, S. (1991) Culture and the self: implications for cognition, emotion and motivation. *Psychological Review*, 98, pp. 224–54.

Marsden, E. (2007) Can educational experiments both test a theory and inform practice? *British Educational Research Journal*, 33 (4), pp. 565–88.

Marshall, C. and Rossman, G. B. (2016) *Designing Qualitative Research* (sixth edition). Thousand Oaks, CA: Sage.

Martin, J. (1999) *Women and the Politics of Schooling in Victorian and Edwardian England.* London and New York: Leicester University Press.

Martin, J. (2010) *Radical Connections: A Journey through Social Histories, Biography and Politics*. London: University of London Institute of Education Press.

Martin, J. (2013) *Making Socialists: Mary Bridges Adams and the Fight for Knowledge and Power, 1855–1939*. Manchester: Manchester University Press.

Martin, J. (2014) Intellectual portraits: politics, professions and identity in twentieth-century England. *History of Education*, 43 (6), pp. 740–67.

Martin, S. (2012) Citizenship, identity and experiential learning in the virtual world. In P. Jerry and L. Lindsey (eds) *Experiential Learning in Virtual Worlds: Opening an Undiscovered Country*. Oxford: Inter-Disciplinary Press, pp. 69–80.

Martin, S. (2013) Exploring identity and citizenship in a virtual world. *International Journal of Virtual and Personal Learning Environments*, 3 (4), pp. 53–70.

Martin, S. (2014) Lessons from the great underground empire: pedagogy, computers and false dawn. In A. Tatnall and B. Davey (eds) *Reflections on the History of Computers in Education*. London: Springer.

Martin, S. (2015) Translational moral constructs from the virtual self. In K. Terry and A. Cheney (eds) *Utilizing Virtual and Personal Learning Environments for Optimal Learning*. Hershey, PA: Information Science Reference, pp. 217–37.

Martin, S. and Vallance, M. (2008) The impact of synchronous inter-networked teacher training in information and communication technology integration. *Computers and Education*, 51 (1), pp. 34–53.

Martin, S., Vallance, M., van Schaik, P. and Wiz, C. (2010) Learning spaces, tasks and metrics for effective communication in Second Life within the context of programming LEGO NXT Mindstrorms TM robots: towards a framework for design and implementation. *Journal of Virtual Worlds Research*, 3 (1), pp. 3–24.

Maslow, A. H. (1954) *Motivation and Personality*. New York: Harper & Row.

Mason, J. (2002) *Qualitative Researching* (second edition). London: Sage.

Masschelein, J. (1991) The relevance of Habermas's communicative turn. *Studies in Philosophy and Education*, 11 (2), pp. 95–111.

Mathison, S. (2007) What is the difference between evaluation and research? And why do we care? In N. L. Smith and P. Brandon (eds) *Fundamental Issues in Evaluation*. New York: Guilford Publishers, pp. 183–96.

Matsumoto, D. and Yoo, S. H. (2006) Toward a new generation of cross-cultural research. *Perspectives on Psychological Science*, 1 (3), pp. 234–50.

Matt, G. E. and Cook, T. D. (2009) Threats to the validity of generalized inferences. In H. M. Cooper, L. V. Hedges and J. C. Valentine (eds) *The Handbook of Research Synthesis and Meta-analysis* (second edition). New York: The Russell Sage Foundation, pp. 537–60.

Matzat, U. and Vrieling, E. M. (2016) Self-regulated learning and social media: a 'natural alliance'? Evidence on students' self-regulation of learning, social media use, and student–teacher relationship. *Learning, Media and Technology*, 41 (1), pp. 73–99.

Mauthner, N. (2012) 'Accounting for our part of the entangled webs we weave': ethical and moral issues in digital sharing. In T. Miller, M. Mauthner, M. Birch and J. Jessop (eds) *Ethics in Qualitative Research* (second edition). London: Sage, pp. 157–76.

Maxwell, J. A. (1992) Understanding and validity in qualitative research. *Harvard Educational Review*, 62 (3), pp. 279–300.

Maxwell, J. A. (2004) Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher*, 33 (2), pp. 3–11.

Maxwell, J. A. (2005) *Qualitative Research Design: An Interactive Approach* (second edition). Thousand Oaks, CA: Sage.

May, D., Wold, K. and Moore, S. (2014) Using interactive online role-playing simulations to develop global competency and to prepare engineering students for a globalised world. *European Journal of Engineering Education*, 40 (5), pp. 522–45.

May, T. (ed.) (2002) *Qualitative Research in Action*. London: Sage.

Mayall, B. (1999) Children and childhood. In S. Hood, B. Mayall and S. Oliver (eds) *Critical Issues in Social Research: Power and Prejudice*. Philadelphia: Open University Press, pp. 10–24.

Mayer-Schönberger, V. and Cukier, K. (2013) *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

Maynard, A. and Chalmers, I. (eds) (1997) *Non-random Reflections on Health Service Research*. London: BMJ Publishing Group.

Mayo, D. (2004) An error-statistical philosophy of evidence. In M. L. Taper and S. R. Lele (eds) *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*. Chicago, IL: University of Chicago Press, pp. 79–118.

Mayring, P. (2004) Qualitative content analysis. In U. Flick, E. von Kardoff and I. Steinke (eds) *A Companion to Qualitative Research*. London: Sage, pp. 266–9.

Meacham, S. J. (1998) Threads of a new language: a response to Eisenhart's 'On the subject of interpretive reviews'. *Review of Educational Research*, 68 (4), pp. 401–7.

Mead, G. H. (1934) *Mind, Self and Society* (ed. Charles Morris). Chicago, IL: University of Chicago Press.

Mears, C. L. (2012) In-depth interviews. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 170–6.

Medawar, P. B. (1972) *The Hope of Progress*. London: Methuen.

Medawar, P. B. (1981) *Advice to a Young Scientist*. London: Pan Books.

Medawar, P. B. (1991) Scientific fraud. In D. Pike (ed.) *The Threat and the Glory: Reflections on Science and Scientists*. Oxford: Oxford University Press, pp. 64–70.

Medd, W. (2002) Complexity and the social world. *International Journal of Social Research Methodology*, 5 (1), pp. 71–81.

Mee, J. F. (1957) 'Supervisory and executive development: a manual for role-playing' – a review. *Industrial and Labor Relations Review*, 11 (1), p. 135.

Mehrens, W. and Kaminski, J. (1989) Methods for improving standardised test scores: fruitful, fruitless or fraudulent? *Educational Measurement: Issues and Practice*, 8 (1), pp. 14–22.

Meinefeld, W. (2004) Hypotheses and prior knowledge in qualitative research. In U. Flick, E. von Kardoff and I. Steinke (eds) *A Companion to Qualitative Research*. London: Sage, pp. 153–8.

Mellor, D. H. (1995) *The Facts of Causation*. London: Routledge.

Melrose, M. J. (1996) Got a philosophical match? Does it matter? In O. Zuber-Skerritt (ed.) *New Directions in Action Research*. London: Falmer, pp. 49–65.

Menter, I. (2013) From interesting times to critical times? Teacher education and educational research in England. *Research in Teacher Education*, 3 (1), pp. 38–40.

Meo, A. I. (2010) Picturing students' habitus: the advantages and limitations of photo-elicitation interviewing in a qualitative study in the city of Buenos Aires. *International Journal of Qualitative Methods*, 9 (2), pp. 149–71.

Mercer, N. (2010) The analysis of classroom talk: methods and methodologies. *British Journal of Educational Psychology*, 80 (1), pp. 1–14.

Merriam, S. B. (1998) *Qualitative Research and Case Study Applications in Education*. San Francisco, CA: Jossey-Bass.

Mertens, D. M. (2007) Transformative paradigm: mixed methods and social justice. *Journal of Mixed Methods Research*, 1 (3), pp. 212–25.

Mertens, D. M. (2012) What comes first? The paradigm or the approach? *Journal of Mixed Methods Research*, 6 (4), pp. 255–7.

Mertens, D. M. and Hesse-Biber, S. (2012) Triangulation and mixed methods research: provocative positions. *Journal of Mixed Methods Research*, 6 (2), pp. 75–9.

Merton, R. K. (1957) *Social Theory and Social Structure* (revised and enlarged edition). New York: The Free Press.

Merton, R. K. (1967) *On Theoretical Sociology: Five Essays Old and New*. New York: The Free Press.

Merton, R. K. (1987) Three fragments from a sociologist's notebooks: establishing the phenomenon, specified ignorance, and strategic research materials. *Annual Review of Sociology*, 13, pp. 1–28.

Merton, R. K. and Kendall, P. L. (1946) The focused interview. *American Journal of Sociology*, 51, pp. 541–57.

Mickelson, R. A. (1994) A feminist approach to researching the powerful in education. In G. Walford (ed.) *Researching the Powerful in Education*. London: UCL Press, pp. 132–50.

Mies, M. (1993) Towards a methodology for feminist research. In M. Hammersley (ed.) *Social Research: Philosophy, Politics and Practice*. London: Sage in association with the Open University Press, pp. 64–82.

Miles, M. B. and Huberman, A. M. (1984) *Qualitative Data Analysis*. Beverly Hills, CA: Sage.

Miles, M. B. and Huberman, A. M. (1994) *Qualitative Data Analysis: An Expanded Sourcebook* (second edition). Thousand Oaks, CA: Sage.

Milgram, S. (1963) Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67 (4), pp. 371–8.

Milgram, S. (1974) *Obedience to Authority*. New York: Harper & Row.

Mill, J. S. (2006) *A System of Logic: Ratiocinative and Inductive, Vol. 7, Books I–III*. Indianapolis, IN: Liberty Fund.

Miller, C. (1995) In-depth interviewing by telephone: some practical considerations. *Evaluation and Research in Education*, 9 (1), pp. 29–38.

Miller, D. and Slater, D. (2000) *The Internet: An Ethnographic Approach.* Oxford: Berg.

Miller, G. and Dingwall, R. (1997) *Context and Method in Qualitative Research*. London: Sage.

Miller, P. V. and Cannell, C. F. (1997) Interviewing for social research. In J. P. Keeves (ed.) *Educational Research, Methodology and Measurement: An International Handbook* (second edition). Oxford: Elsevier Science Ltd, pp. 361–70.

Miller, S., Linn, R. L. and Gronlund, N. E. (2012) *Measurement and Assessment in Teaching* (eleventh edition). New York: Pearson.

Miller, T. and Bell, L. (2002) Consenting to what? Issues of access, gatekeeping and 'informed consent'. In M. Mauthner, M. Birch, J. Jessop and T. Miller (eds) *Ethics in Qualitative Research*. London: Sage, pp. 53–69.

Millmann, J. and Greene, J. (1993) The specification and development of tests of achievement and ability. In R. Linn (ed.) *Educational Measurement* (third edition). Phoenix, AZ: American Council on Education and the Oryx Press, pp. 147–200.

Mills, C. W. (1959) *The Sociological Imagination*. New York: Oxford University Press.

Mills, D. and Morton, M. (2013) *Ethnography in Education*. London: Sage.

Mills, J. (2001) Self-construction through conversation and narrative in interviews. *Educational Review*, 53 (1), pp. 285–301.

Miltiades, H. B. (2008) Interview as a social event: cultural influences experienced while interviewing older adults in India. *International Journal of Social Research Methodology*, 11 (4), pp. 277–91.

Milwain, C. (1998) *Assembling, Maintaining and Disseminating a Social and Educational Controlled Trials Register (SPECTR): A Collaborative Endeavour*. Oxford: UK Cochrane Centre.

Milwain, C., Chalmers, I., Macdonald, S. and Smith, P. (1999) *Cochrane Collaboration Methods Group Newsletter, June*. Available from: www.cochrane-collaboration.com/newslett/MGNews_1999.pdf [Accessed 2 September 2001].

Minnaar, L. and Heystek, J. (2013) Online surveys as data collection instruments in educational research: a feasible option? *South African Journal of Higher Education*, 27 (1), pp. 162–83.

Mishler, E. G. (1986) *Research Interviewing: Context and Narrative*. Cambridge, MA: Harvard University Press.

Mishler, E. G. (1990) Validation in inquiry-guided research: the role of exemplars in narrative studies. *Harvard Educational Review*, 60 (4), pp. 415–42.

Mishler, E. G. (1991) Representing discourse: the rhetoric of transcription. *Journal of Narrative and Life History*, 1 (4), pp. 255–80.

Mitchell, C. (2011) *Doing Visual Research*. London: Sage.

Mitchell, C. (2012) Visual methodologies and social change. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 290–6.

Mitchell, M. and Jolley, J. (1988) *Research Design Explained.* New York: Holt, Rinehart & Winston.

Mitchell, R. G. (1993) *Secrecy and Fieldwork*. London: Sage.

Mitchell, W. and Sloper, T. (2008) *Evaluation of the Pilot Programme of the Integrated Children's System: The Disability Study*. York: University of York Social Policy Research Unit. Available from: www.york.ac.uk/inst/spru/pubs/pdf/ics.pdf [Accessed 2 April 2010].

Mixon, D. (1974) If you won't deceive, what can you do? In N. Armistead (ed.) *Reconstructing Social Psychology*. Harmondsworth: Penguin, pp. 72–85.

Moghaddam, A. (2006) Coding issues in grounded theory. *Issues in Educational Research*, 16 (1), pp. 52–66. Available from: www.iier.org.au/iier16/moghaddam.html [Accessed 28 July 2006].

Monge, D. and Contractor, N. (2003) *Theories of Communication Networks*. Oxford: Oxford University Press.

Monroe, M. C. (2012) Increasing response rates to web-based surveys. *Journal of Extension*, 50 (6), article #6TOT7. Available from: www.joe.org/joe/20-12december/tt7.php?pdf=1 [Accessed 17 March 2016].

Monteiro, E. P. J. F. and Morrison, K. R. B. (2014) Challenges for collaborative blended learning in undergraduate students. *Educational Research and Evaluation*, 20 (7–8), pp. 564–91.

Moon, J., Hossain, D., Sanders, G. L., Garrity, E. J. and Jo, S. (2013) Player commitment to Massively Multiplayer Online Role-Playing Games (MMORPGs): an integrated model. *International Journal of Electronic Commerce*, 17 (4), pp. 7–38.

Moore, L., Graham, A. and Diamond, I. (2003) On the feasibility of conducting randomised trials in education: case study of a sex education intervention. *British Educational Research Journal*, 29 (5), pp. 673–89.

Mora, M. (2010) *Using a Strong Questionnaire to Harvest High-Quality Data.* Available from: www.relevantinsights.com/questionnaire-design [Accessed 21 March 2016].

Mora, M. (2011a) *Why We Need to Avoid Long Surveys*. Available from: www.relevantinsights.com/long-surveys [Accessed 21 March 2016].

Mora, M. (2011b) *Which Rating Scales Should I Use?* Available from: www.relevantinsights.com/rating-scales [Accessed 21 March 2016].

Moreno, J. L. (1939) Psychodramatic shock therapy: a sociometric approach to the problem of mental disorders. *Sociometry*, 2 (1), pp. 1–30.

Morgan, C. (1999) Personal communication with one of the authors. University of Bath, Department of Education.

Morgan, C. (2005) Cultural validity. Personal communication. University of Bath, Department of Education.

Morgan, D. (2007) Paradigm lost and paradigm regained: methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Methods Research*, 1 (2), pp. 212–25.

Morgan, D. L. (1988) *Focus Groups as Qualitative Research*. Beverly Hills, CA: Sage.

Morgan, D. L. (1996) Focus groups. *Annual Review of Sociology*, 22 (1), pp. 129–52.

Morison, M., Moir, J. and Kwansa, T. (2000) Interviewing children for the purposes of primary health care. *Primary Health Care Research and Development*, 1 (2), pp. 113–30.

Moroni, I. (2011) Action research in the library: method, experiences, and a significant case. *Italian Journal of Library and Information Science*, 2 (2), pp. 1–24.

Morris, L. L., Fitz-Gibbon, C. T. and Lindheim, E. (1987) *How to Measure Performance and Use Tests*. Beverly Hills, CA: Sage.

Morris, S. B. (2008) Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, 11 (2), pp. 364–86.

Morrison, K. R. B. (1993) *Planning and Accomplishing School-Centred Evaluation*. Dereham, UK: Peter Francis Publishers.

Morrison, K. R. B. (1995a) Habermas and the school curriculum. Unpublished PhD thesis, School of Education, University of Durham. Durham E-Theses Online. Available from: http://etheses.dur.ac.uk/972 [Accessed 8 July 2004].

Morrison, K. R. B. (1995b) Dewey, Habermas and reflective practice. *Curriculum*, 16 (2), pp. 82–94.

Morrison, K. R. B. (1996a) Developing reflective practice in higher degree students through a learning journal. *Studies in Higher Education*, 21 (3), pp. 317–32.

Morrison, K. R. B. (1996b) Why present school inspections are unethical. *Forum*, 38 (3), pp. 79–80.

Morrison, K. R. B. (1998) *Management Theories for Educational Change*. London: Paul Chapman Publishing.

Morrison, K. R. B. (2001) Randomised controlled trials for evidence-based education: some problems in judging 'what works'. *Evaluation and Research in Education*, 15 (2), pp. 69–83.

Morrison, K. R. B. (2002a) *School Leadership and Complexity Theory*. London: RoutledgeFalmer.

Morrison, K. R. B. (2002b) Education for the open, democratic society in Macau. Paper presented to the Catholic Teachers' Association, Macau, April 2002.

Morrison, K. R. B. (2005) Improving teaching and learning in higher education: metaphors and models for partnership consultancy. *Evaluation and Research in Education*, 17 (1), pp. 31–44.

Morrison, K. R. B. (2006) Sensitive educational research in small states and territories: the case of Macau. *Compare*, 36 (2), pp. 249–64.

Morrison, K. R. B. (2008) Educational philosophy and the challenge of complexity theory. In M. M. Mason (ed.) *Complexity Theory and the Philosophy of Education*. Chichester, UK: John Wiley & Sons, pp. 16–31.

Morrison, K. R. B. (2009) *Causation in Educational Research*. London: Routledge.

Morrison, K. R. B. (2011) Leadership for self-organization: complexity theory and communicative action. *International*

Journal of Complexity in Leadership and Management, 1 (2), pp. 145–63.

Morrison, K. R. B. (2012) Searching for causality in the wrong places. *International Journal of Social Research Methodology*, 15 (1), pp. 15–30.

Morrison, K. R. B. (2013a) Interviewing children in uncomfortable settings: ten lessons for effective practice. *Educational Studies*, 39 (3), pp. 320–37.

Morrison, K. R. B. (2013b) Online and paper evaluations of courses: a literature review and case study. *Educational Research and Evaluation*, 19 (7), pp. 585–604.

Morrison, K. R. B. and Tam, O. I. (2005) Undergraduate students in part-time employment in China. *Educational Studies*, 31 (2), pp. 169–80.

Morrison, K. R. B. and Tang, F. H. (2002) Testing to destruction: a problem in a small state. *Assessment in Education: Principles, Policy and Practice*, 9 (3), pp. 289–317.

Morrison, K. R. B. and Van der Werf, M. (2015) Editorial. *Educational Research and Evaluation*, 21 (3), pp. 185–7.

Morrow, V., Boddy, J. and Lamb, R. (2014) *The Ethics of Secondary Analysis: Learning from the Experience of Sharing Qualitative Data from Young People and Their Families in an International Study of Childhood Poverty.* NOVELLA Working Paper: Narrative Research in Action. London: Thomas Coram Research Unit and the Institute of Education, University of London.

Morse, J. M. (1994) Design in funded qualitative research. In N. Denzin and Y. S. Lincoln (eds) *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage, pp. 220–35.

Morse, J. M., Barrett, M., Mayan, M., Olson, K. and Spiers, J. (2002) Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods*, 1 (2), pp. 1–19.

Moschini, E. (2010) The Second Life Researcher Toolkit: an exploration of inworld tools, methods and approaches for researching educational projects in Second Life. In A. Peachey, J. Gillen, D. Livingstone and S. Smith-Robbins (eds) *Researching Learning in Virtual Worlds*. London: Springer, pp. 31–52.

Moser, C. and Fang, X. W. (2015) Narrative structure and player experience in role-playing games. *International Journal of Human-Computer Interaction*, 31 (2), pp. 146–56.

Moser, C. and Kalton, G. (1977) *Survey Methods in Social Investigation* (second edition). London: Heinemann.

Mostafa, T. (2016) Variation within households in consent to link survey data to administrative records: evidence from the UK Millennium Cohort Study. *International Journal of Social Research Methodology*, 19 (3), pp. 355–75.

Mouly, G. J. (1978) *Educational Research: The Art and Science of Investigation.* Boston, MA: Allyn & Bacon.

Moyles, J. (2002) Observation as a research tool. In M. Coleman and A. J. Briggs (eds) *Research Methods in Educational Leadership*. London: Paul Chapman Publishing, pp. 172–91.

Mueller, C. E. and Hart, C. O. (2010) Effective use of secondary data analysis in gifted education research: opportunities and challenges. *Gifted Children*, 4 (2), pp. 6–11.

Muijs, D. (2004) *Doing Quantitative Research in Education with SPSS.* London: Sage.

Mukherji, P. and Albon, D. (2010) *Research Methods in Early Childhood: An Introductory Guide*. London: Sage.

Munn, N. (1986) *The Fame of Gawa*. Cambridge: Cambridge University Press.

Munn-Giddings, C. (2012) Action research. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 71–5.

Munro, A., Holly, L., Rainbird, H. and Leisten, R. (2004) Power at work: reflections on the research process. *International Journal of Social Research Methodology*, 7 (4), pp. 289–304.

Munshi, F., Lababidi, H. and Alyousef, S. (2015) Low- versus high-fidelity simulations in teaching and assessing clinical skills. *Journal of Taibah University Medical Sciences*, 10 (1), pp. 12–15.

Murphy, M. (ed.) (2013) *Social Theory and Education Research*. London: Routledge.

Murray, F. B. and Raths, J. (1996) Factors in the peer review of reviews. *Review of Educational Research*, 66 (4), pp. 417–21.

National Center for Education Evaluation and Regional Assistance (2003) *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide*. Washington, DC: National Center for Education Evaluation and Regional Assistance.

National Centre for Research Methods (2016) *Practical Exemplars and Survey Analysis*. Avauilable from: www.restore.ac.uk/PEAS/about.php [Accessed 24 March 2016].

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979) *Ethical Principles and Guidelines for the Protection of Human Subjects of Research (The Belmont Report).* Washington, DC: National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.

National Research Council (2002) *Scientific Research in Education* (ed. R. Shavelson and L. Towne, Committee on Scientific Principles for Educational Research). Washington, DC: National Academy Press.

Neal, S. (1995) Researching powerful people from a feminist and anti-racist perspective: a note on gender, collusion and marginality. *British Educational Research Journal*, 21 (4), pp. 517–31.

Neimeyer, G. J. and Hagans, C. L. (2002) More madness in our method? The effects of repertory grid variations on construct differentiation. *Journal of Constructivist Psychology*, 15 (2), pp. 139–60.

Nelson, B. C. and Erlandson, B. E. (2012) *Design for Learning in Virtual Worlds.* New York: Routledge.

Nesfield-Cookson, B. (1987) *William Blake: Prophet of Universal Brotherhood.* London: Crucible.

New South Wales Department of Education and Training (2010) *Action Research in Education*. Sydney, NSW: New South Wales Department of Education and Training, Professional Learning and Leadership Directorate.

Newby, P. (2010) *Research Methods for Education*. Harlow, UK: Pearson Education Ltd.

Ngunjiri, F. W., Hernandez, K. A. C. and Chang, H. W. (2010) Living autoethnography: connecting life and research. *Journal of Research Practice*, 6 (1), pp. 1–17.

Nias, J. (1991) Primary teachers talking: a reflexive account of longitudinal research. In G. Walford (ed.) *Doing Educational Research*. London: Routledge, pp. 147–65.

Nicholson, S. (2015) *The Censorship of British Drama 1900–1968, Vol. 4*. Exeter, UK: University of Exeter Press.

Nicol, R. (2013) Returning to the richness of experience: is autoethnography a useful approach for outdoor educators in promoting pro-environmental behaviour? *Journal of Adventure Education and Outdoor Learning*, 31 (1), pp. 3–17.

Niemeyer, R., Johnson, A. and Monroe, A. E. (2014) Role play for classroom management: providing a lodestar for alternative-route teachers. *The Educational Forum*, 78 (3), pp. 338–46.

Nisbet, J. and Watt, J. (1984) Case study. In J. Bell, T. Bush, A. Fox, J. Goodey and S. Goulding (eds) *Conducting Small-Scale Investigations in Educational Management*. London: Harper & Row, pp. 79–92.

Nisbett, R. E. (2005) *The Geography of Thought*. London: Nicholas Brealey Publishers.

Noblit, G. W. and Hare, R. D. (1988) *Meta-ethnography: Synthesizing Qualitative Studies*. Newbury Park, CA: Sage.

Noffke, S. E. and Zeichner, K. M. (1987) Action research and teacher thinking. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Nolen, A. L. and Vander Putten, J. (2007) Action research in education: addressing gaps in ethical principles and practices. *Educational Researcher*, 36 (7), pp. 401–7. Available from: www.aera.net/uploadedFiles/Publications/Journals/Educational_Researcher/3607/10EDR07_401-407.pdf [Accessed 24 April 2010].

Norris, N. (1990) *Understanding Educational Evaluation*. London: Kogan Page.

Norton, D. F. and Norton, M. J. (eds) (2000) *David Hume: A Treatise of Human Nature*. Oxford: Oxford University Press.

Nóvoa, A. (2000) Ways of saying, ways of seeing. Public images of teachers (19th and 20th century). *Paedagogica Historica*, 36 (10), pp. 21–51.

Noy, C. (2008) Sampling knowledge: the hermeneutics of snowball sampling in qualitative research. *International Journal of Social Research Methodology*, 11 (4), pp. 327–44.

Nuttall, D. (1987) The validity of assessments. *European Journal of Psychology of Education*, 11 (2), pp. 109–18.

Oakley, A. (1981) Interviewing women: a contradiction in terms. In H. Roberts (ed.) *Doing Feminist Research*. London: Routledge & Kegan Paul, pp. 30–61.

Oakley, A. (1998) Gender, methodology and people's ways of knowing. *Sociology*, 34 (4), pp. 707–31.

Oakley, A. (1999) Paradigm wars: some thoughts on a personal and public trajectory. *International Journal of Social Research Methodology*, 2 (3), pp. 247–54.

O'Connell, A. A. and McCoach, D. B. (eds) (2008) *Multilevel Modeling of Educational Data*. Charlotte, NC: Information Age Publishing.

O'Donoghue, D. (2010) Classrooms as installations: a conceptual framework for analysing classroom photographs from the past. *History of Education*, 39 (3), pp. 401–15.

OECD (2012) Good practices in survey design step-by-step. In *Measuring Regulatory Performance: A Practitioner's Guide to Perception Surveys*. Paris: OECD. Available from: http://dx.doi.org/10.1787/9789262167179-6-en [Accessed 20 March 2016].

Ogawa, R. T. and Malen, B. (1991) Towards rigor in reviews of multivocal literatures: applying the exploratory case study method. *Review of Educational Research*, 61 (3), pp. 265–86.

Ohm, P. (2009) Broken promises of privacy: responding to the surprising failure of anonymization. *UCLA Law Review*, 57 (6), p. 1701.

Oja, S. N. and Smulyan, L. (1989) *Collaborative Action Research: A Developmental Approach.* Lewes, UK: Falmer.

Oldroyd, D. (1986) *The Arch of Knowledge: An Introductory Study of the History of the Philosophy and Methodology of Science.* New York: Methuen.

O'Leary, C., Santos Sánchez, D. and Thompson, M. (eds) (2015) *Global Insights on Theatre Censorship*. London: Routledge.

Oliver, P. (2003) *The Student's Guide to Research Ethics*. Maidenhead, UK: Open University Press.

O'Neill, C. (1995) *Drama Worlds: A Framework for Process Drama*. Portsmouth, NH: Heinemann.

O'Neill, C. (ed.) (2014) *Dorothy Heathcote on Education and Drama: Essential Writings*. London: Routledge.

Onwuegbuzie, A. and Johnson, R. B. (2006) The validity issue in mixed research. *Research in the Schools*, 13 (1), pp. 48–63.

Onwuegbuzie, A. J. and Leech, N. L. (2005) On becoming a pragmatic researcher: the importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology*, 8 (5), pp. 375–87.

Onwuegbuzie, A. J. and Leech, N. L. (2006a) Linking research methods to mixed methods data analysis procedures. *The Qualitative Report*, 11 (3), pp. 474–98.

Onwuegbuzie, A. J. and Leech, N. L. (2006b) Validity and qualitative research: an oxymoron? *Quality and Quantity*, 41 (2), pp. 233–49.

Onwuegbuzie, A. J. and Leech, N. L. (2007) Sampling designs in qualitative research: making the sampling process more public. *The Qualitative Report*, 12 (2), pp. 238–54.

Oppenheim, A. N. (1992) *Questionnaire Design, Interviewing and Attitude Measurement*. London: Pinter Publishers Ltd.

Orton-Johnson, K. (2010) Ethics in online research: evaluating the ESRC Framework for Research Ethics categorisation of risk. *Sociological Research Online*, 15 (4). Available from: www.socresonline.org.uk/15/4/13.html [Accessed 6 April 2016].

Orwin, R. G. and Vevea, J. L. (2009) Evaluating coding decisions. In H. M. Cooper, L. V. Hedges and J. C. Valentine (eds) *The Handbook of Research Synthesis and Meta-analysis* (second edition). New York: The Russell Sage Foundation, pp. 177–203.

Osborn, D. and Costas, L. (2013) Role-playing in counsellor student development. *Journal of Creativity in Mental Health*, 8 (1), pp. 92–103.

Osgood, C. E., Suci, G. S. and Tannenbaum, P. H. (1957) *The Measurement of Meaning*. Urbana, IL: University of Illinois.

O'Sullivan, C. (2015) The day that Shrek was almost rescued: doing process drama with children with an autism spectrum disorder. In P. Duffy (ed.) *What Was I Thinking: A Reflective Practitioner's Guide to (Mis)Adventures in Drama Education*. New York: Peter Lang, pp. 219–42.

O'Sullivan, C. (2016a) *Practical Guide to Planning Drama in Education*. Beijing: Renmin University Press.

O'Sullivan, C. (2016b) Building bridges: drama in the social education of young people with Autism Spectrum Disorder. The World Festival of Theatre for Young Audiences, ASSITEJ (International Association of Theatre for Children and Young People) Artistic Gathering, Birmingham Repertory Theatre, Birmingham, UK, 2–9 July.

O'Sullivan, C., McKernan, D., O'Halloran, S. and Rowland, J. (2009) *Asperger Syndrome: A Practical Guide for Parents, Teachers, Young People and Other Professionals* [two-hour DVD]. Dublin, Ireland: Specialist AV Ltd.

O'Toole, J. and Haseman, B. (1992) *Dramawise: An Introduction to GCSE Drama*. Oxford: Heinemann Educational Publishers.

Ovadia, S. (2004) Ratings and rankings: reconsidering the structure of values and their measurement. *International Journal of Social Research Methodology*, 7 (5), pp. 403–14.

Owen, N., Fox, A. and Bird, T. (2016) The development of a small scale survey instrument of UK teachers to study professional use (and non-use) of and attitudes to social media. *International Journal of Research and Method in Education*, 39 (2), pp. 170–93.

Oxenham, J. (ed.) (1984) *Education versus Qualifications?* London: George Allen & Unwin.

Paccagnella, L. (1997) Getting the seat of your pants dirty: strategies for ethnographic research on virtual communities. *Journal of Computer Mediated Communication*, 3 (1). DOI: 10.1111/j.1083-6101.1997.tb00065.x.

Pallant, J. (2016) *SPSS Survival Manual* (sixth edition). Maidenhead, UK: Open University Press.

Pampaka, M., Hutcheson, G. and Williams, J. (2016) Handling missing data: analysis of a challenging data set using multiple imputation. *International Journal of Research and Method in Education*, 39 (1), pp. 19–37.

Papacharissi, Z. (ed.) (2011) *A Networked Self*. London: Routledge.

Parker, H. J. (1974) *View from the Boys*. Newton Abbot, UK: David & Charles.

Parker, L. and Lynn, M. (2002) What's race got to do with it? Critical race theory's conflicts with and connections to qualitative research methodology and epistemology. *Qualitative Inquiry*, 8 (1), pp. 7–22.

Parlett, M. and Hamilton, D. (1976) Evaluation as illumination. In D. Tawney (ed.) *Curriculum Evaluation Today: Trends and Implications*. London: Macmillan, pp. 84–101.

Paterson, B. L., Thorne, S. E., Canam, C. and Jillings, C. (2001) *Meta-study of Qualitative Health Research: A Practical Guide to Meta-analysis and Meta-synthesis*. Thousand Oaks, CA: Sage.

Paterson, L. and Goldstein, H. (1991) New statistical methods for analysing social structures: an introduction of multilevel models. *British Educational Research Journal*, 17 (4), pp. 387–93.

Patrick, J. (1973) *A Glasgow Gang Observed*. London: Methuen.

Patton, M. Q. (1980) *Qualitative Evaluation Methods*. Beverly Hills, CA: Sage.

Patton, M. Q. (1990) *Qualitative Evaluation and Research Methods* (second edition). London: Sage.

Patton, M. Q. (1998) Research vs evaluation. 16 January Message board post quoted in S. Mathison (2007) What is the difference between evaluation and research? And why do we care? In N. L. Smith and P. Brandon (eds) *Fundamental Issues in Evaluation*. New York: Guilford Publishers, pp. 183–96.

Patton, M. Q. (2002) *Qualitative Research and Evaluation Methods* (third edition). Thousand Oaks, CA: Sage.

Patton, M. Q. (2014) *Evaluation Flash Cards: Embedding Evaluative Thinking in Organizational Culture*. St. Paul, MN: Otto Bremer Foundation.

Paugh, P. and Robinson, E. (2011) Keeping a 'vigilant critique': unpacking critical praxis as teacher educators. *International Journal of Qualitative Studies in Education*, 24 (3), pp. 363–78.

Paul, J. A., Baker, H. M. and Cochran, J. D. (2012) Effect of online social networking on student academic performance. *Computers in Human Behavior*, 28 (6), pp. 2117–27.

Paulus, T. and Lester, J. N. (2016) ATLAS.ti for conversation and discourse analysis studies. *International Journal of Social Research Methodology*, 19 (4), pp. 405–28.

Paulus, T., Woods, M., Atkins, D. P. and Macklin, R. (2017) The discourse of QDAS: reporting practices of ATLAS.ti and NVivo users with implications for best practices. *International Journal of Social Research Methodology*, 20 (1), pp. 35–47.

Pawson, R. (2006) *Evidence-Based Policy: A Realist Perspective*. London: Sage.

Pawson, R. (2008) Middle range theory and programme theory evaluation: from provenance to practice. In F. Leeuw and J. Vassan (eds) *Mind the Gap: Perspectives on Policy Evaluation and the Social Sciences*. New Brunswick, NJ: Transaction Press, pp. 171–202.

Pawson, R. (2013) *The Science of Evaluation: A Realist Manifesto*. London: Sage.

Pawson, R., Greenhalgh, T., Harvey, G. and Walshe, K. (2005) Realist review: a new method of systematic review designed for complex policy interventions. *Journal of Health Services Research and Policy*, 10 (1), pp. 21–34.

Payne, G., Dingwall, R., Payne, J. and Carter, M. (1980) *Sociology and Social Research*. London: Routledge & Kegan Paul.

Peachey, A., Gillen, J., Livingstone, D. and Smith-Robbins, S. (eds) (2010) *Researching Learning in Virtual Worlds.* London: Springer.

Pearce, G., Thøgersen-Ntoumani, C. and Duda, J. L. (2014) The development of synchronous text-based instant messaging as an online interviewing tool. *International Journal of Social Research Methodology*, 17 (6), pp. 677–92.

Pearce, W. and Raman, S. (2014) The new randomized controlled trials (RCT) movement in public policy: challenges of epistemic governance. *Policy Sciences*, 47 (4), pp. 387–402.

Pearl, J. (2009) *Causality: Models, Reasoning and Inference* (second edition). Cambridge: Cambridge University Press.

Pearson, G. (2009) The researcher as hooligan: where 'participant' observation means breaking the law. *International Journal of Social Research Methodology*, 12 (3), pp. 243–55.

Pelias, R. J. (2015) A story located in 'should': toward a productive feature for qualitative inquiry. *Qualitative Inquiry*, 21 (7), pp. 609–11.

Petticrew, M. and Roberts, H. (2006) *Systematic Reviews in the Social Sciences: A Practical Guide*. Malden, MA: Blackwell.

Phelps, R. and Graham, A. (2010) Exploring the complementarities between complexity and action research: the story of *Technology Together*. *Cambridge Journal of Education*, 40 (2), pp. 183–97.

Phillips, D. C. and Burbules, N. C. (2000) *Postpositivism and Educational Research.* Lanham, MD: Rowman & Littlefield Publishers.

Phillips, R. (1998) Some methodological and ethical dilemmas in élite-based research. *British Educational Research Journal*, 24 (1), pp. 5–20.

Piaget, J. (1932) *The Moral Judgement of the Child*. London: Routledge & Kegan Paul.

Piggot-Irvine, E., Rowe, W. and Ferkins, L. (2015) Conceptualizing indicator-domains for evaluating action research. *Educational Action Research*, 23 (4), pp. 545–66.

Pilliner, A. (1973) *Experiment in Educational Research* (Course E341). Milton Keynes, UK: Open University Press.

Pillow, W. S. (2010) Dangerous reflexivity: rigour, responsibility and reflexivity in qualitative research. In P. Thompson and M. Walker (eds) *The Routledge Doctoral Student's Companion*. New York: Routledge, pp. 270–82.

Pink, D. H. (2011) *Drive: The Surprising Truth about What Motivates Us*. New York: Riverhead Books.

Pink, S. (2007) *Doing Visual Ethnography* (second edition). London: Sage.

Pinney, C. (2004) *'Photos of the Gods': The Printed Image and Political Struggle in India*. London: Reaktion Books.

Pinto, M. (2000) *Doing Research with People*. New Delhi: Participatory Research in Asia.

Pitcher, E. N. (2016) Analysing whispers: college students' representation and reproduction of sociocultural discourses about bodies, relationships, and (hetero)sexuality using a mobile application. *International Journal of Qualitative Studies in Education*, 29 (5), pp. 714–30.

Pituch, K. A. and Stevens, J. P. (2016) *Applied Mulitivariate Statistics for the Social Sciences* (sixth edition). London: Routledge.

Plewis, I. (1997) *Statistics in Education*. London: Arnold.

Plewis, I. and Mason, P. (2005) What works and why: combining quantitative and qualitative approaches in large-scale evaluations. *International Journal of Social Research Methodology*, 8 (3), pp. 185–94.

Plummer, K. (1983) *Documents of Life: An Introduction to the Problems and Literature of a Humanistic Method*. London: Allen & Unwin.

Plummer, K. (1995) *Telling Sexual Stories: Power, Change and Social Worlds*. London: Routledge.

Plummer, K. (2001) *Documents of Life, Vol. 2: An Invitation to a Critical Humanism*. London: Sage.

Polkinghorne, D. E. (1995) Narrative configuration in qualitative analysis. *International Journal of Qualitative Studies in Education*, 8 (1), pp. 5–23.

Polkinghorne, D. E. (2007) Validity issues in narrative research. *Qualitative Inquiry*, 13 (4), pp. 471–86.

Pope, C., Mays, N. and Popay, J. (2007) *Synthesizing Qualitative and Quantitative Health Evidence: A Guide to Methods*. Maidenhead, UK: Open University Press.

Popper, K. (1968) *The Logic of Scientific Discovery* (second edition). London: Hutchinson.

Popper, K. (1980) *Conjectures and Refutations* (third edition). London: Routledge & Kegan Paul.

Potter, J. and Wetherell, M. (1994) Analysing discourse. In A. Brymer and R. G. Burgess (eds) *Analysing Qualitative Data*. London: Routledge, pp. 46–66.

Powell, M. (2007) The importance of using open-ended questions when interviewing children. *La Gazette: Une Publication de la Gendarmerie Royale du Canada*, 69 (2), pp. 26–7.

Powney, J. and Watts, M. (1987) *Interviewing in Educational Research*. London: Routledge & Kegan Paul.

Preissle, J. (2006) Envisioning qualitative inquiry: a view across four decades. *International Journal of Qualitative Studies in Education*, 19 (6), pp. 685–95.

Priede, C., Jokinen, A., Ruuskanen, E. and Farrall, S. (2014) Which probes are most useful when undertaking cognitive interviews? *International Journal of Social Research Methodology*, 17 (5), pp. 559–68.

Priest, S. (2001) A program evaluation primer. *Journal of Experiential Education*, 24 (1), pp. 34–40.

Pring, R. (1984) The problems of confidentiality. In M. Skilbeck (ed.) *Evaluating the Curriculum in the Eighties*. Sevenoaks, UK: Hodder & Stoughton, pp. 38–44.

Pring, R. (2015) *Philosophy of Educational Research* (third edition). London: Bloomsbury Academic.

Prosser, J. and Burke, C. (2011) Image-based educational research: childlike perspectives. *LEARNing Landscapes*, 4 (2), pp. 257–73.

Prosser, J. and Loxley, A. (2008) *Introducing Visual Methods*. NCRM Methodological Review. Manchester: ESRC National Centre for Research Methods. Available from: www.ncrm.ac.uk/research/outputs/publications [Accessed 17 May 2010].

Prosser, J., Clark, A. and Wiles, R. (2008) *Visual Research Ethics at the Crossroads*. Working Paper #10. Manchester: ESRC National Centre for Research Methods. Available from: www.socialsciences.manchester.ac.uk/realities/publications/workingpapers/10-2008-11-realities-prosseretal.pdf [Accessed 17 May 2010].

Punch, K. F. (2003) *Survey Research: The Basics*. London: Sage.

Punch, K. F. (2005) *Introduction to Social Research: Quantitative and Qualitative Approaches* (second edition). London: Sage.

Pyle, A. (2013) Engaging young children in research through photo elicitation. *Early Child Development and Care*, 183 (11), pp. 1544–58.

Quantz, R. A. (1992) On critical ethnography (with some postmodern considerations). In M. LeCompte, W. L. Millroy and J. Preissle (eds) *The Handbook of Qualitative Research in Education*. London: Academic Press Ltd, pp. 447–506.

Radford, M. (2006) Researching classrooms: complexity and chaos. *British Educational Research Journal*, 32 (2), pp. 177–90.

Radford, M. (2007) Action research and the challenge of complexity. *Cambridge Journal of Education*, 37 (2), pp. 263–78.

Radford, M. (2008) Prediction, control and the challenge of complexity. *Oxford Review of Education*, 34 (5), pp. 505–20.

Raento, M., Oulasvirta, A. and Eagle, N. (2009) Smartphones. *Sociological Methods and Research*, 37 (3), pp. 426–54.

Raffe, D., Bundell, I. and Bibby, J. (1989) Ethics and tactics: issues arising from an educational survey. In R. G. Burgess (ed.) *The Ethics of Educational Research*. Lewes, UK: Falmer, pp. 13–30.

Raghunathan, T. (2015) *Missing Data Analysis in Practice*. New York: Chapman & Hall/CRC.

Ragin, C. C. (1987) *The Comparative Method: Moving beyond Qualitative and Quantitative Strategies*. Berkeley, CA: University of California Press.

Ragin, C. C. (1992) Introduction: cases of 'what is a case?'. In C. C. Ragin and H. S. Becker (eds) *What Is a Case?* Cambridge: Cambridge University Press, pp. 1–18.

Ragin, C. C. (1994a) *Constructing Social Research*. Thousand Oaks, CA: Sage.

Ragin, C. C. (1994b) Introduction to Qualitative Comparative Analysis. In T. Janoski and A. Hicks (eds) *The Comparative Political Economy of the Welfare State*. Cambridge: Cambridge University Press, pp. 299–319.

Ragin, C. C. (2000) *Fuzzy-Set Social Science*. Chicago, IL: University of Chicago Press.

Ragin, C. C. (2004) Combining qualitative and quantitative research. Paper presented at the National Science Foundation Workshop on Scientific Foundations of Qualitative Research in Arlington, VA. Available from: www.nsf.gov/pubs/2004/nsf04219/nsf04219_6.pdf [Accessed 3 August 2106].

Ragin, C. C. (2006a) The limitations of net-effects thinking. In B. Rihoux and H. Grimm (eds) *Innovative Comparative Methods for Policy Analysis*. New York: Springer, pp. 13–41.

Ragin, C. C. (2006b) Set relations in social research: evaluating their consistency and coverage. *Political Analysis*, 14 (3), pp. 291–310.

Ragin, C. C. (2008) *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago, IL: University of Chicago Press.

Ragin, C. C. and Davey, S. (2014) *fs/QCA* [computer program], Version 2.5. Irvine, CA: University of California. Available from: www.socsci.uci.edu/~cragin/fsQCA/software.shtml [Accessed 4 August 2016].

Ragin, C. C. and Schneider, G. A. (2011) Case-oriented theory building and theory testing. In M. Williams and W. P. Vogt (eds) *The SAGE Handbook of Innovation in Social Research Methods*. London: Sage, pp. 150–66.

Ramírez, G. B. and Palu-ay, L. (2015) 'You don't look like your profile picture': the ethical implications of researching online identities in higher education. *Educational Research and Evaluation*, 21 (2), pp. 139–53.

Rao, D. and Stupans, I. (2012) Exploring the potential of role play in higher education: development of a typology and teacher guidelines. *Innovations in Education and Teaching International*, 49 (4), pp. 427–36.

Rasmussen, D. M. (1990) *Reading Habermas*. Oxford: Basil Blackwell Ltd.

Raudenbush, A. and Bryk, A. (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage.

Ravenscroft, A. and McAllister, S. (2006) Digital games, learning in cyberspace: a dialogical approach. *E-Learning*, 3 (1), pp. 38–51.

Raynes-Goldie, K. (2010) Aliases, creeping and wall cleaning: understanding privacy in the age of Facebook. *First Monday*, 15 (1).

Raynes-Goldie, K. (2012) *Privacy in the Age of Facebook: Discourse, Architecture, Consequences*. PhD thesis, Curtin University, Perth, Australia.

Reams, P. and Twale, D. (2008) The promise of mixed methods: discovering conflicting realities in the data. *International Journal of Research and Method in Education*, 31 (2), pp. 133–42.

Redline, C. D., Dillman, D. A., Carley-Baxter, L. and Creecy, R. (2002) Factors that influence reading and comprehension in self-administered questionnaires. Paper presented at the Workshop on Item-Nonresponse and Data Quality, Basel, Switzerland, 10 October.

Reed-Danahay, D. E. (1997) *Auto/ethnography: Rewriting the Self and the Social*. Oxford: Berg.

Rees, R. and Oliver, S. (2012) Stakeholder perspectives and participation in reviews. In D. Gough, S. Oliver and J. Thomas (eds) *An Introduction to Systematic Reviews*. London: Sage, pp. 17–34.

Reichardt, C. S. and Rallis, S. F. (1994) Qualitative and quantitative inquiries are not incompatible: a call for a new partnership. In C. S. Reichardt and S. F. Rallis (eds) *The Qualitative–Quantitative Debate: New Perspectives*. San Francisco, CA: Jossey-Bass, pp. 85–92.

Reichenbach, H. (1956) *The Direction of Time*. Berkeley and Los Angeles: University of California Press.

Reid, W. A. and Holley, B. J. (1972) An application of repertory grid techniques to the study of choice of university. *British Journal of Educational Psychology*, 42 (1), pp. 52–9.

Reinharz, S. T. (1979) *On Becoming a Social Scientist: From Survey Research and Participant Obseration to Experiential Analysis*. San Francisco, CA: Jossey-Bass.

Reinking, D. and Bradley, B. A. (2008) *Formative and Design Experiments: Approaches to Language and Literacy Research*. New York: Teachers College Press.

Reips, U.-D. (2002a) Internet-based psychological experimenting: five dos and don'ts. *Social Science Computer Review*, 20 (3), pp. 241–9.

Reips, U.-D. (2002b) Standards for Internet-based experimenting. *Experimental Psychology*, 49 (4), pp. 243–56.

Reips, U.-D. (2009) The methodology of internet-based experiments. In H. Joinson, K. McKenna, T. Postmes and U.-D. Reips (eds) *The Oxford Handbook of Internet Psychology*. Oxford: Oxford University Press, pp. 373–90.

Renkema, J. (2004) *Introduction to Discourse Studies*. Amsterdam: John Benjamins Publishing.

Renzetti, C. M. and Lee, R. M. (1993) *Researching Sensitive Topics*. London: Sage.

Rex, J. (1974) *Approaches to Sociology: An Introduction to Major Trends in British Sociology*. London: Routledge & Kegan Paul.

Reynolds, C. R. and Kamphaus, R. W. (eds) (2003) *Handbook of Psychological and Educational Assessment of Children: Intelligence, Aptitude and Achievement* (second edition). New York: Guilford Press.

Reynolds, R. and de Zwart, M. (2010) The duty to 'play': ethics, EULAs and MMOs. *International Journal of Internet Research Ethics*, 3 (1), pp. 48–68.

Reynolds, T. J. and Gutman, J. (1988) Laddering theory, method, analysis, and interpretation. *Journal of Advertising Research*, 28 (1), pp. 11–31.

Rheingold, H. (2000) *The Virtual Community: Homesteading on the Electronic Frontier*. Cambridge, MA: MIT Press.

Rice, J. M. (1897) The futility of the spelling grind. Cited in G. de Landsheere (1997) History of educational research. In J. P. Keeves (ed.) *Educational Research, Methodology, and Measurement: An International Handbook* (second edition). Oxford: Elsevier Science Ltd, pp. 8–16.

Richard, V. M. and Lahman, M. K. E. (2015) Photo-elicitation: reflexivity on method, analysis, and graphic portraits. *International Journal of Research and Method in Education*, 38 (1), pp. 3–22.

Richards, L. (2002) Qualitative computing: a methods revolution? *International Journal of Social Research Methodology*, 5 (3), pp. 263–76.

Richardson, L. (2000) Writing: a method of inquiry. In N. K. Reichardt and S. F. Rallis (eds) *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage, pp. 923–48.

Richardson, L. (2001) Getting personal: writing-stories. *Qualitative Studies in Education*, 14 (1), pp. 33–8.

Ricoy, M. and Feliz, T. (2016) Twitter as a learning community in higher education. *Educational Technology & Society*, 19 (1), pp. 237–48.

Ridley, D. (2010) *The Literature Review: A Step-by-Step Guide for Students* (second edition). London: Sage.

Riessman, C. K. (1993) *Narrative Analysis.* Newbury Park, CA: Sage.

Riessman, C. K. (2008) *Narrative Methods for the Human Sciences*. London: Sage.

Riley, M. W. (1963) *Sociological Research, Vol. 1: A Case Approach*. New York: Harcourt, Brace and World Inc.

Ringer, F. (1997) *Max Weber's Methodology*. Cambridge, MA: Harvard University Press.

Riordan, C. M. and Vandenburg, R. J. (1994) A central question in cross-cultural research: do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management*, 20 (3), pp. 643–71.

Risjord, M. (2014) *Philosophy of Social Science*. London: Routledge.

Ritchie, S. M. and Rigano, D. L. (2007) Solidarity through collaborative research. *International Journal of Qualitative Studies in Education*, 20 (2), pp. 129–50.

Riva, G., Lorerti, P., Lunghi, M., Vatalaro, F. and Davide, F. (2003) Presence 2010: the emergence of ambient intelligence. In G. Riva, F. Davide and W. A. Ijsselsteijn (eds) *Being There: Concepts, Effects and Measurements of User Presence in Synthetic Environments*. Amsterdam: International Operations Press.

Rivers, A., Wickramasekera II, I. E., Pekala, R. J. and Rivers, J. A. (2016) Empathetic features and absorption in fantasy role-playing. *American Journal of Clinical Hypnosis*, 58 (3), pp. 286–94.

Roberts, J. T. (2004) There are no laws of the social sciences. In C. Hitchcock (ed.) *Contemporary Debates in Philosophy of Science*. Oxford: Blackwell Publishing, pp. 151–67.

Roberts, L. D. and Allen, P. J. (2015) Exploring ethical issues associated with using online surveys in educational research. *Educational Research and Evaluation*, 21 (2), pp. 95–108.

Roberts, S. (1991) Exploring alternative paradigms in higher education. Paper presented at the Annual Meeting of the Association for the Study of Higher Education, Boston, MA (ERIC Document Reproduction Service No. ED 339 327).

Robinson, B. (1982) *Tutoring by Telephone: A Handbook*. Milton Keynes, UK: Open University Press.

Robrecht, L. (1995) Grounded theory: evolving methods. *Qualitative Health Research*, 5 (2), pp. 169–77.

Robson, C. (1993) *Real World Research*. Oxford: Blackwell.

Robson, C. (2002) *Real World Research* (second edition). Oxford: Blackwell.

Robson, K. and Pevalin, D. (2016) *Multilevel Modeling in Plain Language*. London: Sage.

Robson, S. (2014) The Analysing Children's Creative Thinking framework: development of an observation-led approach to identifying and analysing young children's creative thinking. *British Educational Research Journal*, 40 (1), pp. 121–34.

Roderick, R. (1986) *Habermas and the Foundations of Critical Theory*. Basingstoke, UK: Macmillan.

Rodrigues, D. and Rodrigues, R. (2000) *The Research Paper and the World Wide Web*. Upper Saddle River, NJ: Prentice-Hall.

Rogan, A. I. and de Kock, D. M. (2005) Chronicles from the classroom: making sense of the methodology and methods of narrative analysis. *Qualitative Inquiry*, 11 (4), pp. 628–49.

Rogers, C. R. (1942) *Counselling and Psychotherapy*. Boston, MA: Houghton Mifflin.

Rogers, C. R. (1945) The non-directive method as a technique for social research. *American Journal of Sociology*, 50, pp. 279–83.

Rogers, C. R. (1969) *Freedom to Learn*. Columbus, OH: Merrill Pub. Co.

Rohlfing, I. and Starke, P. (2013) Building on solid ground: robust case selection in multi-method research. *Swiss Political Science Review*, 19 (4), pp. 492–512.

Rohner, R. and Katz, L. (1970) Testing for validity and reliability in cross-cultural research. *American Anthropologist*, 72, pp. 1068–73.

Roman, L. G. and Apple, M. (1990) Is naturalism a move away from positivism? Materialist and feminist approaches to subjectivity in ethnographic research. In E. Eisner and A. Peshkin (eds) *Qualitative Inquiry in Education: The Continuing Debate*. New York: Teachers College Press, pp. 38–73.

Roos, L. (2011) From holy groves to holy ghost. *Studia Theologica – Nordic Journal of Theology*, 65 (1), pp. 54–73.

Rose, D. and Sullivan, O. (1993) *Introducing Data Analysis for Social Scientists*. Buckingham, UK: Open University Press.

Rose, G. (2007) *Visual Methodologies* (second edition). London: Sage.

Rose, J. (2001) *The Intellectual Life of the British Working Classes*. New Haven, CT: Yale University Press.

Rosenberg, A. (2010) Virtual world research ethics and the private/public distinction. *International Journal of Internet Research Ethics*, 3 (12), pp. 23–37.

Rosenthal, R. (1980) Combining probabilities and the file drawer problem. *Evaluation in Education*, 4, pp. 18–21.

Rosiek, J. and Atkinson, B. (2007) The inevitability and importance of genres in narrative research on teaching practice. *Qualitative Inquiry*, 13 (4), pp. 499–521.

Rosier, M. J. (1997) Survey research methods. In J. P. Keeves (ed.) *Educational Research, Methodology and Measurement: An International Handbook* (second edition). Oxford: Elsevier Science Ltd, pp. 154–62.

Ross, K. N. and Rust, K. (1997) Sampling in survey research. In J. P. Keeves (ed.) *Educational Research, Methodology, and Measurement: An International Handbook* (second edition). Oxford: Elsevier Science Ltd, pp. 427–38.

Roszak, T. (1970) *The Making of a Counter Culture*. London: Faber & Faber.

Roszak, T. (1972) *Where the Wasteland Ends*. London: Faber & Faber.

Roth, W. D. and Mehta, J. D. (2002) The Rashomon effect. *Sociological Methods and Research*, 31 (2), pp. 131–73.

Roth, W. F. (2009) Auto/ethnography and the question of ethics. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 10 (1), pp. 1–10.

Rothstein, H. R., College, B., Turner, H. M. and Lavenberg, J. G. (2004) *The Campbell Collaboration Information Retrieval Policy Brief*. Available from: www.campbellcollaboration.org/MG/IRMGPolicyBriefRevised.pdf [Accessed 5 June 2015].

Rowan-Kenyon, H. T., Martínez Alemán, A. M., Gin, K., Blakeley, B., Gismondi, A., Lewis, J., McCready, A., Zepp, D. and Knight, S. (2016) Social media in higher education. *ASHE Higher Education Report*, 42 (5).

Rowbotham, S. (1975) *Hidden from History*. London: Pluto Press.

Roztocki, N. and Lahri, N. A. (2002) Is the applicability of web-based surveys for academic research limited to the field of information technology? Proceedings of the 36th Hawaii International Conference on System Sciences. Available from: http://csdl.computer.org/comp/proceedings/hicss/2003/1874/08/187480262a.pdf [Accessed 26 February 2005].

Ruane, J. M. (2005) *Essentials of Research Methods: A Guide to Social Science Research*. Oxford: Blackwell.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, 63 (3), pp. 581–92.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Ruddin, L. P. (2006) You can generalize stupid! Social scientists, Bent Flyvberg, and case study methodology. *Qualitative Inquiry*, 12 (4), pp. 797–812.

Ruddock, J. (1981) *Evaluation: A Consideration of Principles and Methods*. Manchester: University of Manchester.

Rudé, G. (1964) *The Crowd in History*. New York: Wiley & Sons.

Ruel, E. E., Wagner, W. E. and Gillespie, B. J. (2015) *The Practice of Survey Research: Theory and Applications*. Thousand Oaks, CA: Sage.

Ruspini, E. (2002) *Introduction to Longitudinal Research*. London: Routledge.

Russell, B. (1959) *The Problems of Philosophy*. Oxford: Oxford University Press.

Rutkowski, L., Gonzalez, E., Joncas, M. and von Davier, M. (2010) International large-scale assessment data: issues in secondary analysis and reporting. *Educational Researcher*, 39 (2), pp. 142–51.

Rybas, N. and Gajjala, R. (2007) Developing cyberethnographic research methods for understanding digitally mediated identities. *Forum: Qualitative Social Research*, 8 (3). Available from: www.qualitative-research.net/index.php/fqs/article/view/282/620. [Accessed 20 April 2010].

Sacks, H. (1992) *Lectures on Conversation* (ed. G. Jefferson). Oxford: Basil Blackwell.

Sadowski, W. and Stanney, K. (2002) Presence in virtual environments. In K. Stanney (ed.) *Handbook of Virtual Environments: Design, Implementation, and Applications*. Mahwah, NJ: Erlbaum, pp. 791–806.

Said, E. (1978) *Orientalism*. London: Routledge & Kegan Paul.

Saklofske, D. H., Andrews, J. J. W., Janzen, H. L. and Phye, G. D. (2001) *Handbook of Psychoeducational Assessment: A Practical Handbook*. New York: Academic Press.

Salmon, G., Ross, B., Pechenkina, E. and Chase, A. (2015) The space for social media in structured online learning. *Research in Learning Technology*, 23 (1), pp. 1–14.

Salmon, P. (1976) Grid measures with child subjects. In P. Slater (ed.) *The Measurement of Intrapersonal Space by Grid Technique, Vol. 1: Explorations of Interpersonal Space*. London: Wiley, pp. 15–46.

Salmon, W. C. (1998) *Causality and Explanation*. Oxford: Oxford University Press.

Sampson, D. G., Ifenthaler, D., Spector, J. M. and Isaias, P. (eds) (2014) *Digital Systems for Open Access to Formal and Informal Learning*. New York: Springer.

Samuel, R. (ed.) (1975) *Village Life and Labour*. History Workshop Series. London: Routledge & Kegan Paul.

Sanday, A. (1993) The relationship between educational research and evaluation and the role of the local education authorities. In R. G. Burgess (ed.) *Educational Research and Evaluation for Policy and Practice*. London: Falmer, pp. 32–43.

Sandelowski, M. (2001) Real qualitative researchers do not count: the use of numbers in qualitative research. *Research in Nursing and Health*, 24 (3), pp. 230–40.

Sandelowski, M. and Barroso, J. (2002) Reading qualitative studies. *International Journal of Qualitative Methods*, 1 (1), pp. 74–108.

Sandelowski, M. and Barroso, J. (2007) *Handbook for Synthesizing Qualitative Research*. New York: Springer.

Sandelowski, M., Voils, C. I. and Knafl, G. (2009) On quantitizing. *Journal of Mixed Methods Research*, 3 (3), pp. 208–22.

Santonus, M. (1998) *Simple, Yet Complex*. Available from: www.cio.com/archive/enterprise/041598_qanda_content.html [Accessed 10 November 2000].

Sapsford, R. (1999) *Survey Research*. London: Sage.

Sartre, J. P. (1964) *The Words*. New York: Braziller.

Sartre, J. P. (1976) *Sartre on Theatre*. New York: Random House.

Saunders, P. (1997) Social mobility in Britain: an empirical evaluation of two competing explanations. *Sociology*, 31 (2), pp. 261–88.

Schaefer, D. R. and Dillman, D. A. (1998) Development of a standard e-mail methodology: results of an experiment. *Public Opinion Quarterly*, 62 (3), pp. 378–97.

Schatzman, L. and Strauss, A. L. (1973) *Field Research: Strategies for a Natural Sociology*. Englewood Cliffs, NJ: Prentice-Hall Inc.

Schein, E. (1992) *Organizational Culture and Leadership* (second edition). San Francisco, CA: Jossey-Bass.

Schellenberg, E. G. (2004) Music lessons enhance IQ. *Psychological Science*, 15 (8), pp. 511–54.

Schensul, S. L., Schensul, J. J. and LeCompte, M. D. (1999) *Essential Ethnographic Methods: Observations, Interviews and Questionnaires*. Walnut Creek, CA: AltaMira Press.

Scheper-Hughes, N. (1979) *Saints, Scholars and Schizophrenics: Mental Illness in Rural Ireland*. Berkeley, CA: University of California Press.

Scheurich, J. J. (1995) A postmodernist critique of research interviewing. *Qualitative Studies in Education*, 8 (3), pp. 239–52.

Scheurich, J. J. (1996) The masks of validity: a deconstructive investigation. *International Journal of Qualitative Studies in Education*, 9 (1), pp. 49–60.

Scheurich, J. J. (1997) *Research Method in the Postmodern*. London: Falmer.

Schindler, L. (2009) The production of 'vis-ability': an ethnographic video analysis of a martial arts class. In U. Kissmann (ed.) *Video Interaction Analysis*. Frankfurt: Peter Lang, pp. 135–54.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H. and Shavelson, R. J. (2007) *Estimating Causal Effects Using Experimental and Observational Designs*. Washington, DC: American Educational Research Association.

Schneider, C. Q. and Rohlfing, I. (2013) Combining QCA and process tracing in set-theoretic multi-method research. *Sociological Methods and Research*, 42 (4), pp. 559–97.

Schneider, C. Q. and Rohlfing, I. (2016) Case studies nested in fuzzy-set QCA on sufficiency: formalizing case selection and causal inference. *Sociological Methods and Research*, 45 (3), pp. 526–68.

Schneider, C. Q. and Wagemann, C. (2012) *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. Cambridge: Cambridge University Press.

Schneider, C. Q. and Wagemann, C. (2015) Assessing ESA on what it is designed for: a reply to Cooper and Glaesser. *Field Methods*. DOI: 10.1177/1525822X15598977.

Schnurr, M. A., De Santo, E. M. and Green, A. D. (2014) What do students learn from a role-play simulation of an international negotiation? *Journal of Geography in Higher Education*, 38 (3), pp. 401–14.

Schnurr, M. A., De Santo, E. M., Green, A. D. and Taylor, A. (2015) Investigating student perceptions of knowledge acquisition within a role-play simulation of the Convention on Biological Diversity. *Journal of Geography*, 114 (3), pp. 94–107.

Schofield, W. (1996) Survey sampling. In R. Sapsford and V. Jupp (eds) *Data Collection and Analysis*. London: Sage and the Open University Press, pp. 25–55.

Schön, D. (1983) *The Reflective Practitioner: How Professionals Think in Action*. London: Temple Smith.

Schön, D. (1987) *Educating the Reflective Practitioner*. San Francisco, CA: Jossey-Bass.

Schonlau, M., Van Soest, A., Kapteyn, A. and Couper, M. (2009) Selection bias in web surveys and the use of propensity scores. *Sociological Methods and Research*, 37 (3), pp. 291–318.

Schuemie, M. J., Van der Straaten, P. and Van der Mast, C. A. (2001) Research on presence in VR: a survey. *Cyberpsychology and Behavior*, 4 (2), pp. 183–201.

Schumacker, R. A. and Lomax, R. G. (2005) *A Beginner's Guide to Structural Equation Modeling* (second edition). Mahwah, NJ: Lawrence Erlbaum Associates.

Schutz, A. (1962) *Collected Papers*. The Hague: Nijhoff.

Schwandt, T. A. (1998) The interpretive review of educational matters: is there any other kind? *Review of Educational Research*, 68 (4), pp. 409–12.

Schwartz, N. and Bienias, J. (1990) What mediates the impact of response alternatives on frequency reports of mundane behaviors? *Applied Cognitive Psychology*, 4 (1), pp. 61–72.

Schwartz, N., Grayson, C. A. and Knauper, B. (1998) Formal meaning of rating scales and the interpretation of questions. *International Journal of Public Opinion Research*, 10 (2), pp. 177–83.

Schwartz, N., Knauper, B., Rippler, H. J., Noelle-Neumann, E. and Clark, F. (1991) Rating scales: numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, 55 (4), pp. 570–82.

Schwarz, S. and Reips, U.-D. (2001) CGI versus Javascript: a web experiment on the reversed hindsight bias. In U.-D. Reips and M. Bosnjak (eds) *Dimensions of Internet Science*. Lengerich, Germany: Pabst Science, pp. 75–90.

Scott, J. (1990) *A Matter of Record: Documentary Sources in Social Research*. Cambridge: Polity Press.

Scott, J. W. (1986) Gender: a useful category of historical analysis. *American Historical Review*, 91 (5), pp. 1053–75.

Scott, S. (1985) Feminist research and qualitative methods: a discussion of some of the issues. In R. G. Burgess (ed.) *Issues in Educational Research: Qualitative Methods*. Lewes, UK: Falmer, pp. 67–85.

Scriven, M. (1991) *Evaluation Thesaurus*. Newbury Park, CA: Sage.

Scriven, M. (2004) Michael Scriven on the differences between evaluation and social science research. *The Evaluation Exchange*, 9 (1), p. 4.

Searle, J. (1969) *Speech Acts*. London: Cambridge University Press.

Sears, R., Maccoby, E. and Levin, H. (1957) *Patterns of Child Rearing*. Palo Alto, CA: Stanford University Press.

Sears, R. R., Rau, L. and Alpert, R. (1965) *Identification and Child Rearing*. Stanford, CA: Stanford University Press.

Seawright, J. and Gerring, J. (2008) Case selection techniques in case study research: a menu of qualitative and quantitative options. *Political Research Quarterly*, 61 (2), pp. 294–308.

Sechrest, L. and Sidana, S. (1995) Quantitative and qualitative methods: is there an alternative? *Evaluation and Program Planning*, 18 (1), pp. 77–87.

Seedhouse, D. (1998) *Ethics: The Heart of Healthcare*. Chichester, UK: Wiley.

Seel, N. M. (2011) Design experiments. In N. Seel (ed.) *Encyclopedia of the Sciences of Learning*. New York: Springer, pp. 925–8.

Segall, A. (2001) Critical ethnography and the invocation of voice: from the field/in the field – single exposure, double

standard? *Qualitative Studies in Education*, 14 (4), pp. 579–92.
Seidel, J. and Kelle, U. (1995) Different functions of coding in the analysis of textual data. In U. Kelle (ed.) *Computer-Aided Qualitative Data Analysis: Theory, Methods and Practice*. London: Sage, pp. 52–61.
Seidman, I. E. (1998) *Interviewing as Qualitative Research* (second edition). New York: Teachers College Press.
Sellers, S. C. (2002) Testing theory through teaching theatrics. *Journal of Nursing Education*, 41 (11), pp. 498–500.
Sellitz, C., Wrightsman, L. S. and Cook, S. W. (1976) *Research Methods in Social Relations*. New York: Holt, Rinehart & Winston.
Seltzer, M. and Rickles, J. (2012) Multilevel analysis. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 356–67.
Selwyn, N. (2010) Looking beyond learning: notes towards the critical study of educational technology. *Journal of Computer Assisted Learning*, 26 (1), pp. 65–73.
Selwyn, N. and Stirling, E. (2016) Social media and education … now the dust has settled. *Learning, Media and Technology*, 41 (1), pp. 1–5.
Sequeira, G. M., Howroid, S., MacPherson, S. and Lo, O. Y. (1996) *The Poverty Research Project*. Hong Kong: City University of Hong Kong.
Serafini, A. (ed.) (1989) *Ethics and Social Concern*. New York: Paragon House.
Shadish, W. R. (n.d.) Links to meta-analysis software. Available from: http://faculty.ucmerced.edu/wshadish/Meta-Analysis%20Software.htm [Accessed 4 July 2004].
Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin Company.
Shapiro, B. L. (1990) A collaborative approach to help novice science teachers reflect on changes in their construction of the role of the science teacher. *Alberta Journal of Educational Research*, 36 (3), pp. 203–22.
Shapiro, S. and Leopold, L. (2012) A critical role for role-playing pedagogy. *TESL Canada Journal*, 29 (2), pp. 120–30.
Shaughnessy, J. J., Zechmeister, E. B. and Zechmeister, J. S. (2003) *Research Methods in Psychology* (sixth edition). New York: McGraw-Hill.
Shavelson, R. J. and Berliner, D. C. (1991) Erosion of the education research infrastructure: a reply to Finn. In D. S. Anderson and B. J. Biddle (eds) *Knowledge for Policy: Improving Education through Research*. London: Falmer, pp. 79–84.
Shavelson, R. J., Phillips, D. C., Towne, L. and Feuer, M. J. (2003) On the science of education design studies. *Educational Researcher*, 32 (1), pp. 25–8.
Shaw, C. M. (2010) Designing and using simulations and role-play exercises. *The International Studies Encyclopedia*. Available from: www.isacompendium.com/subscriber/tocnode?id=g9781444336597_yr2011_chunk_g97814443365976_ss1-4 [Accessed 20 April 2013].
Shaw, E. L. (1992) The influence of methods instruction on the beliefs of preservice elementary and secondary science teachers: preliminary comparative analyses. *School Science and Mathematics*, 92 (1), pp. 14–22.
Sheehy, K. (2010) Virtual environments: issues and opportunities for researching inclusive educational practices. In A. Peachey, J. Gillen, D. Livingstone and S. Smith-Robbins (eds) *Researching Learning in Virtual Worlds*. London: Springer, pp. 1–16.
Sheffield Hallam University (2016) *Can Randomised Controlled Trials Revolutionise Educational Research?* Sheffield, UK: Sheffield Institute of Education, Sheffield Hallam University. Available from: www4.shu.ac.uk/research/ceir/randomised-controlled-trials-1 [Accessed 10 March 2016].
Sheridan, T. B. (1992) Musings on telepresence and virtual presence. *Presence: Teleoperators and Virtual Environments*, 1 (1), pp. 120–5.
Shipman, G. (1964) Role playing in the class. *Improving College and University Teaching*, 12 (1), pp. 21–3.
Sholle, D. (1992) Authority on the left: critical pedagogy, postmodernism and vital strategies. *Cultural Studies*, 6 (2), pp. 271–89.
Shropshire, K. O., Hawdon, J. E. and Witte, J. C. (2009) Web survey design: balancing measurement, response, and topical interest. *Sociological Methods and Research*, 37 (3), pp. 344–70.
Shulman, A., Joost, H., Kavanagh, R., Ratner, B., Tooley, T., Jarecki, A. and Smerling, M. (2011) *Catfish*. Universal City, CA: Universal Studios Home Entertainment.
Shuy, R. W. (2003) In-person versus telephone interviewing. In J. A. Holstein and J. F. Gubrium (eds) *Inside Interviewing: New Lenses, New Concerns*. Thousand Oaks, CA: Sage, pp. 195–3.
Sieber, J. E. (1992) *Planning Ethically Responsible Research: A Guide for Students and Internal Review Boards*. Beverly Hills, CA: Sage.
Sieber, J. E. and Stanley, B. (1988) Ethical and professional dimensions of socially sensitive research. *American Psychologist*, 43 (1), pp. 49–55.
Siegel, S. (1956) *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.
Sikes, P. (2006) On dodgy ground? Problematics and ethics in educational research. *International Journal of Research and Method in Education*, 29 (1), pp. 105–17.
Silverman, D. (1985) *Qualitative Methodology and Sociology: Describing the Social World*. Brookfield, VT: Gower.
Silverman, D. (1993) *Interpreting Qualitative Data*. London: Sage.
Simon, B. (1965) *Education and the Labour Movement, 1870–1920*. London: Lawrence and Wishart.
Simon, J. L. (1978) *Basic Research Methods in Social Science*. New York: Random House.
Simon, M. and Tierney, R. D. (2011) Use of vignettes in educational research on sensitive teaching functions such as assessment. Paper presented at the 24th International Congress for School Effectiveness and Improvement, Limassol, Cyprus, January.
Simon, M. K. (2011) *Developing Research Questions*. Seattle, WA: Dissertation Success, LLC. Available from: http://dissertationrecipes.com/wp-content/uploads/2011/04/Developing-Research-Questions.pdf [Accessed 14 September 2015].
Simons, H. (1982) Conversation piece: the practice of interviewing in case study research. In R. McCormick (ed.)

Calling Education to Account. London: Heinemann, pp. 239–46.
Simons, H. (1989) *Getting to Know School in a Democracy*. London: Falmer.
Simons, H. (2000) Damned if you do, damned if you don't: ethical and political dilemmas in education. In H. Simons and R. Usher (eds) *Situated Ethics in Educational Research*. London: RoutledgeFalmer, pp. 39–55.
Simons, H. (2009) *Case Study Research in Practice*. London: Sage.
Simons, H. (2015) Interpret in context: generalizing from the single case in evaluation. *Evaluation*, 21 (2), pp. 173–88.
Simons, H. and Usher, R. (eds) (2000) *Situated Ethics in Educational Research*. London: RoutledgeFalmer.
Simpson, M. and Tuson, J. (2003) *Using Observations in Small-Scale Research: A Beginner's Guide* (revised edition). Glasgow: University of Glasgow, the SCRE Centre.
Sinclair, J. M. and Coulthard, M. (1975) *Towards an Analysis of Discourse*. Oxford: Oxford University Press.
Sirin, S. R. (2005) Socioeconomic status and academic achievement: a meta-analytic review of research. *Review of Educational Research*, 75 (3), pp. 417–53.
Skåreus, E. (2009) Pictorial analysis in research on education: methods and concepts. *International Journal of Research and Method in Education*, 32 (2), pp. 167–83.
Skelton, C., Francis, B. and Smulyan, L. (2006) *The Sage Handbook of Gender and Education*. London: Sage.
Skinner, Q. (2002) *Visions of Politics, Vol. 1: Regarding Method*. Cambridge: Cambridge University Press.
Slater, M. and Steed, A. (2007) A virtual presence counter. *Presence Teleoperators and Virtual Environments*, 9 (5), pp. 413–34.
Slavin, R. E. (2008) Perspectives on evidence-based research in education: what works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37 (1), pp. 5–14.
Slavin, R. E., Chamberlain, A., Daniels, C. and Madden, N. A. (2009) The Reading Edge: a randomized evaluation of a middle school cooperative reading program. *Effective Education*, 1 (1), pp. 13–26.
Sloane, F. C. and Gorard, S. (2003) Exploring modeling aspects of design experiments. *Educational Researcher*, 32 (1), pp. 29–31.
Small, M. L. (2011) How to conduct a mixed methods study: recent trends in a rapidly growing literature. *Annual Review of Sociology*, 37, pp. 55–84.
Smeyers, P. and Verhesschen, P. (2001) Narrative analysis as philosophical research: bridging the gap between the empirical and the conceptual. *International Journal of Qualitative Studies in Education*, 14 (1), pp. 71–84.
Smith, B. G. (1998) *The Gender of History: Men, Women and Historical Practice*. Cambridge, MA: Harvard University Press.
Smith, E. (2008) *Using Secondary Data in Educational and Social Research*. Maidenhead, UK: Open University Press.
Smith, E. (2011) *Using Numeric Secondary Data in Education Research*. British Educational Research Association online resource. Available from: www.bera.ac.uk/wp-content/uploads/2014/03/Using-numeric.pdf?noredirect=1 [Accessed 2 March 2016].
Smith, E. (2012) Secondary data. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 125–30.
Smith, H. A. and Haslett, S. J. (2016) Design and implementation issues in surveying the views of young children in ethnolinguistically diverse developing country contexts. *International Journal of Research and Method in Education*, 39 (2), pp. 131–50.
Smith, H. W. (1975) *Strategies of Social Research: The Methodological Imagination*. London: Prentice-Hall.
Smith, H. W. (1991) *Strategies of Social Research* (third edition). Orlando, FL: Holt, Rinehart & Winston.
Smith, M. (2013) Evidence-based education: is it really that straightforward? *Guardian*, 26 March. Available from: www.theguardian.com/teacher-network/2013/mar/26/teachers-research-evidence-based-education [Accessed 10 June 2016].
Smith, M. L. and Glass, G. V. (1987) *Research and Evaluation in Education and the Social Sciences*. Englewood Cliffs, NJ: Prentice-Hall.
Smith, O. (2016) Integrating role-play with case study and carbon footprint monitoring: a transformative approach to enhancing learners' social behavior for a more sustainable environment. *International Journal of Environmental & Science Education*, 11 (6), pp. 1323–35.
Smithson, J. (2000) Using and analysing focus groups: limitations and possibilities. *International Journal of Social Research Methodology*, 3 (2), pp. 103–19.
Smyth, J. (1989) Developing and sustaining critical reflection in teacher education. *Journal of Teacher Education*, 40 (2), pp. 2–9.
Smyth, J. and Hattam, R. (2000) Intellectual as hustler: researching against the grain of the market. *British Educational Research Journal*, 26 (2), pp. 157–75.
Smyth, J. D., Dillman, D. A., Christian, L. M. and Stern, M. J. (2004) How visual grouping influences answers to Internet surveys. Paper presented at the American Association for Public Opinion Research, Phoenix.
Snell, J. (2011) Interrogating video data: systematic quantitative analysis versus micro-ethnographic analysis. *International Journal of Social Research Methodology*, 14 (3), pp. 253–8.
Snijders, T. A. B. and Bosker, R. J. (2012) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (second edition). London: Sage.
Social Research Association (2003) *Ethical Guidelines*. Available from: www.the-sra.org.uk/ethics03.pdf [Accessed 15 May 2005].
Solberg, A. (2014) Reflections on interviewing children living in difficult circumstances: courage, caution and co-production. *International Journal of Social Research Methodology*, 17 (3), pp. 233–48.
Solomon, D. J. (2001) *Conducting Web-Based Surveys ERIC Digest*. ED458291. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. Available from: www.ericdigests.org/2002-2/surveys.htm [Accessed 14 April 2004].
Solove, D. J. (2004) *The Digital Person: Technology and Privacy in the Information Age*. New York: New York University Press.
Solove, D. J. (2006) A taxonomy of privacy. *University of Pennsylvania Law Review*, 154 (3), pp. 477–560.

Somekh, B. (1995) The contribution of action research to development in social endeavours: a position paper on action research methodology. *British Educational Research Journal*, 21 (3), pp. 339–55.

Somekh, B. (2006) *Action Research: A Methodology for Change and Development*. Maidenhead, UK: Open University Press.

Souto-Manning, M. (2014) Critical narrative analysis: the interplay of critical discourse and narrative analysis. *International Journal of Qualitative Studies in Education*, 27 (2), pp. 159–80.

Sparkes, A. C. (2000) Autoethnography and narratives of self: reflections on criteria in action. *Sociology of Sport Journal*, 17 (1), pp. 21–43.

Spector, P. E. (1993) Research designs. In M. L. Lewis-Beck (ed.) *Experimental Design and Methods: International Handbook of Quantitative Applications in the Social Sciences, Vol. 3*. London: Sage, pp. 1–74.

Spindler, G. and Spindler, L. (1992) Cultural process and ethnography: an anthropological perspective. In M. LeCompte, W. L. Millroy and J. Preissle (eds) *The Handbook of Qualitative Research in Education*. London: Academic Press Ltd, pp. 53–92.

Spirtes, P., Glymour, C. and Scheines, R. (2001) C*ausation, Prediction, and Search* (second edition). Cambridge, MA: MIT Press.

Spradley, J. P. (1979) *The Ethnographic Interview*. New York: Holt, Rinehart & Winston.

Spradley, J. P. (1980) *Participant Observation*. New York: Holt, Rinehart & Winston.

St Pierre, E. A. and Jackson, A. Y. (2014) Qualitative data analysis after coding. *Qualitative Inquiry*, 20 (6), pp. 715–19.

St Pierre, E. A. and Roulston, K. (2006) The state of qualitative inquiry: a contested science. *International Journal of Qualitative Studies in Education*, 19 (6), pp. 673–84.

Stables, A. (1990) Differences between pupils from mixed and single-sex schools in their enjoyment of school subjects and in their attitude to Science in school. *Educational Review*, 42 (3), pp. 221–30.

Stacey, J. (1988) Can there be a feminist ethnography? *Women's Studies International Forum*, 11 (1), pp. 21–7.

Stacey, R. D. (1992) *Managing the Unknowable*. San Francisco, CA: Jossey-Bass.

Stacey, R. D. (2000) *Strategic Management and Organisational Dynamics* (third edition). Harlow, UK: Pearson Education Limited.

Stake, R. E. (1994) Case studies. In N. K. Denzin and Y. S. Lincoln (eds) *Handbook of Qualitative Research*. London: Sage, pp. 236–47.

Stake, R. E. (1995) *The Art of Case Study Research*. Thousand Oaks, CA: Sage.

Stake, R. E. (2005) Qualitative case studies. In N. Denzin and Y. Lincoln (eds) *The Sage Handbook of Qualitative Research* (third edition). Thousand Oaks, CA: Sage, pp. 443–66.

Starr, L. J. (2010) The use of autoethnography in educational research: locating who we are and what we do. *Canadian Journal for New Scholars in Education*, 3 (1), pp. 1–9.

Stenbacka, C. (2001) Qualitative research requires quality concepts of its own. *Management Decision*, 39 (7), pp. 551–5.

Stenhouse, L. (1975) *An Introduction to Curriculum Research and Development*. London: Heinemann.

Stenhouse, L. (1979) *What Is Action Research?* (mimeo). Norwich, UK: Classroom Action Research Network.

Stenhouse, L. (1985) Case study methods. In T. Husen and T. N. Postlethwaite (eds) *International Encyclopaedia of Education* (first edition). Oxford: Pergamon, pp. 640–6.

Stevens, G., O'Donnell, V. L. and Williams, L. (2015) Public domain or private data? Developing an ethical approach to social media research in an inter-disciplinary project. *Educational Research and Evaluation*, 21 (2), pp. 154–67.

Stevens, R. (2015) Role-play and student engagement: reflections from the classroom. *Teaching in Higher Education*, 20 (5), pp. 481–92.

Stewart, J. and Yalonis, C. (2001) *Internet-Based Surveys and Sampling Issues*. Communique Partners. Available from: www.communiquepartners.com/white_papers/sampling_issues_and_the_internet_briefing_paper.pdf [Accessed 26 January 2005].

Stewart, M. (2001) *The Co-evolving Organization*. Rutland, UK: Decomplexity Associates Ltd. Available from: www.decomplexity.com/Coevolving%20Organization%20VU.pdf [Accessed 14 November 2001].

Stiggins, R. J. (2001) *Student-Involved Classroom Assessment* (third edition). Upper Saddle River, NJ: Merrill Prentice-Hall.

Stillar, G. F. (1998) *Analysing Everyday Texts: Discourses, Rhetoric and Social Perspectives*. London: Sage.

Stirling, E. (2016) Technology, time and transition in higher education: two different realities of everyday Facebook use in the first year of university in the UK. *Learning, Media and Technology*, 41 (1), pp. 100–18.

Stock, W. A. (1994) Systematic coding for research synthesis. In H. M. Cooper and L. V. Hedges (eds) *The Handbook of Research Synthesis*. New York: The Russell Sage Foundation, pp. 125–38.

Stokoe, E. (2013) The (in)authenticity of simulated talk: comparing role-played and actual interaction and the implications for communication training. *Research on Language and Social Interaction*, 46 (2), pp. 165–85.

Stokoe, E. (2014) The Conversation Analytic Role-play Model (CARM): a method for training communication skills as an alternative to simulated role-play. *Research on Language and Social Interaction*, 47 (3), pp. 255–65.

Strange, V., Forest, S., Oakley, A. and The Ripple Study Team (2003) Using research questionnaires with young people in schools: the influence of social context. *International Journal of Social Research Methodology*, 6 (4), pp. 337–46.

Strauss, A. L. (1987) *Qualitative Analysis for Social Scientists*. Cambridge: Cambridge University Press.

Strauss, A. L. and Corbin, J. (1990) *Basics of Qualitative Research*. Newbury Park, CA: Sage.

Strauss, A. L. and Corbin, J. (1994) Grounded theory methodology: an overview. In N. Denzin and Y. Lincoln (eds) *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage, pp. 273–85.

Strauss, A. L. and Corbin, J. (1998) *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (second edition). Thousand Oaks, CA: Sage.

Strauss, A. L. and Corbin, J. (2008) *Basics of Qualitative Research: Grounded Theory Procedures and Techniques* (third edition). Newbury Park, CA: Sage.

Strohmetz, D. B. and Skleder, A. A. (1992) The use of role-play in teaching research ethics: a validation study. *Teaching of Psychology*, 19 (2), pp. 106–8.

Stronach, I. and Morris, B. (1994) Polemical notes on educational evaluation in an age of 'policy hysteria'. *Evaluation and Research in Education*, 8 (1–2), pp. 5–19.

Stuchbury, K. and Fox, A. (2009) Ethics in educational research: introducing a methodological tool for effective ethical analysis. *Cambridge Journal of Education*, 39 (4), pp. 489–504.

Stufflebeam, D. L. (1967) The use and abuse of evaluation in Title III. *Theory into Practice*, 6 (3), pp. 126–33.

Stufflebeam, D. L. (2001) *Evaluation Models: A New Direction for Evaluation*. New York: Jossey-Bass.

Sturman, A. (1997) Case study methods. In J. P. Keeves (ed.) *Educational Research, Methodology and Measurement: An International Handbook* (second edition). Oxford: Elsevier Science, pp. 61–6.

Sturman, A. (1999) Case study methods. In J. P. Keeves and G. Lakomski (eds) *Issues in Educational Research*. Oxford: Elsevier Science Ltd, pp. 103–12.

Stylianou, S. (2008) Interview control questions. *International Journal of Social Research Methodology*, 11 (3), pp. 239–56.

Sudman, S. and Bradburn, N. M. (1982) *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco, CA: Jossey-Bass.

Sullivan, G. M. (2011) Getting off the 'gold standard': randomized controlled trials and education research. *Journal of Graduate Medical Education*, 3 (3), pp. 285–98.

Sumathipala, A. and Murray, J. (2006) New approach to translating instruments for cross-cultural research: a combined qualitative and quantitative approach for translation and consensus generation. *International Journal of Methods in Psychiatric Research*, 9 (2), pp. 87–95.

Suri, H. (1999) The process of synthesising qualitative research: a case-study. Paper presented at the biennial international conference of the Association for Qualitative Research (AQR) held in Melbourne, Australia, 6–10 July.

Suri, H. (2011) Purposeful sampling in qualitative research synthesis. *Qualitative Research Journal*, 11 (2), pp. 63–75.

Suri, H. (2013) Epistemological pluralism in qualitative research synthesis. *International Journal of Qualitative Studies in Education*, 26 (7), pp. 889–911.

Suri, H. (2014) *Towards Methodologically Inclusive Research Synthesis*. London: Routledge.

Suter, L. E. (2005) Multiple methods: research methods in education projects at NSF. *International Journal of Research and Method in Education*, 28 (2), pp. 171–81.

Suto, S. M. I. and Nádas, R. (2009) Why are some GCSE examination questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features. *Research Papers in Education*, 24 (3), pp. 335–77.

Sutton, A. J. (2009) Publication bias. In H. M. Cooper, L. V. Hedges and J. C. Valentine (eds) *The Handbook of Research Synthesis and Meta-analysis* (second edition). New York: The Russell Sage Foundation, pp. 435–52.

Sveinsdottir, T. (2008) Virtual identity as practice: exploring the relationship between role-players and their characters in the massively multiplayer online game *Star Wars Galaxies*. PhD thesis submitted to Department of Sociology, University of Surrey. Available from: http://epubs.surrey.ac.uk/2112 [Accessed 22 September 2016].

Swain, J. (2006) An ethnographic approach to researching children in junior school. *International Journal of Social Research Methodology*, 9 (3), pp. 199–213.

Swain, J., Heyman, B. and Gillman, M. (1998) Public research, private concerns: ethical issues in the use of open-ended interviews with people who have learning difficulties. *Disability and Society*, 13 (1), pp. 21–36.

Swantz, M. (1996) A personal position paper on participatory research: personal quest for living knowledge. *Qualitative Inquiry*, 2 (1), pp. 120–36.

Tabachnick, B. G. and Fidell, L. S. (2013) *Using Multivariate Statistics* (sixth edition). Harlow, UK: Pearson Education Ltd.

Tananuraksakul, N. (2014) Use of Facebook group as blended learning and learning management system in writing. *Teaching English with Technology*, 14 (3), pp. 3–15.

Tandon, R. (2005a) Introduction: revisiting the roots. In R. Tandon (ed.) *Participatory Research: Revisiting the Roots*. New Delhi: Mosaic Books, pp. vii–xiii.

Tandon, R. (2005b) A critique of monopolistic research. In R. Tandon (ed.) *Participatory Research: Revisiting the Roots*. New Delhi: Mosaic Books, pp. 3–8.

Tandon, R. (2005c) Participatory research: main concepts and issues. In R. Tandon (ed.) *Participatory Research: Revisiting the Roots*. New Delhi: Mosaic Books, pp. 22–39.

Tandon, R. (2005d) Knowledge as power. In R. Tandon (ed.) *Participatory Research: Revisiting the Roots*. New Delhi: Mosaic Books, pp. 40–53.

Tandon, R. (2005e) Dialogue. In R. Tandon (ed.) *Participatory Research: Revisiting the Roots*. New Delhi: Mosaic Books, pp. 275–94.

Tashakkori, A. and Creswell, J. W. (2007) Exploring the nature of research questions in mixed methods research. *Journal of Mixed Methods Research*, 1 (3), pp. 207–11.

Tashakkori, A. and Teddlie, C. (eds) (2003) *Handbook of Mixed Methods Research.* Thousand Oaks, CA: Sage.

Task Group on Assessment and Testing (1988) *National Curriculum: Testing and Assessment – A Report*. London: HMSO.

Taylor, C. and Gibbs, G. R. (2010) *What Is Qualitative Data Analysis (QDA)?* Available from: http://onlineqda.hud.ac.uk/Intro_QDA/what_is_qda.php [Accessed 9 June 2016].

Taylor, T. L. (2009) The assemblage of play. *Games and Culture*, 4 (4), pp. 331–9.

Tedder, M. (2012) Biographical research methods. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 322–9.

Teddlie, C. and Tashakkori, A. (2006) A general typology of research designs featuring mixed methods. *Research in the Schools*, 13 (1), pp. 12–28.

Teddlie, C. and Tashakkori, A. (2009) *Foundations of Mixed Methods Research*. Thousand Oaks, CA: Sage.

Teddlie, C. and Yu, F. (2007) Mixed methods sampling: a typology with examples. *Journal of Mixed Methods Research*, 1 (1), pp. 77–100.

Tesch, R. (1990) *Qualitative Research: Analysis Types and Software*. London: Falmer.

Tess, P. A. (2013) The role of social media in higher education classes (real and virtual): a literature review. *Computers in Human Behavior*, 29 (5), pp. A60–A68.

Teusner, A. (2016) Insider research, validity issues, and the OHS professional: one person's journey. *International Journal of Social Research Methodology*, 19 (1), pp. 85–96.

Thapar-Björket, S. and Henry, M. (2004) Reassessing the research relationship: location, position and power in fieldwork accounts. *International Journal of Social Research Methodology*, 7 (5), pp. 363–81.

Thiem, A. (2015) Standards of good practice and the methodology of necessary conditions in Qualitative Comparative Analysis: a critical view on Schneider and Wagemann's theory-guided/enhanced standard analysis. Available from: www.compasss.org/wpseries/Thiem2015.pdf [Accessed 12 August 2016].

Thiem, A. (2016) *QCApro: Professional Functionality for Performing and Evaluating Qualitative Comparative Analysis*. R Package Version 1.1–1. Available from: www.alrik-thiem.net/software [Accessed 4 August 2016].

Thiem, A., Baumgartner, M. and Bol, D. (2016) Still lost in translation! A correction of three misunderstandings between configurational comparativists and regressional analysts. *Comparative Political Studies*, 49 (6), pp. 742–74.

Thiem, A., Spöhel, R. and Duşa, A. (2016) Enhancing sensitivity diagnostics for qualitative comparative analysis: a combinatorial approach. *Political Analysis*, 24 (1), pp. 104–20.

Thissen, D. (1990) Reliability and measurement precision. In H. Wainer (ed.) *Computer Adaptive Testing: A Primer*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 161–86.

Thomas, D. R. (2006) A general inductive approach for analysing qualitative evaluation data. *American Journal of Education*, 27 (2), pp. 237–46.

Thomas, G. (2010) Doing case study: abduction, not induction, phronesis not theory. *Qualitative Inquiry*, 16 (7), pp. 575–82.

Thomas, G. (2011) A typology for the case study in social science following a definition, discourse and structure. *Qualitative Inquiry*, 17 (6), pp. 511–21.

Thomas, G. and James, D. (2006) Reinventing grounded theory: some questions about theory, ground and discovery. *British Educational Research Journal*, 32 (6), pp. 767–95.

Thomas, G. and Myers, K. (2015) *The Anatomy of the Case Study*. London: Sage.

Thomas, J. (1993) *Doing Critical Ethnography*. Newbury Park, CA: Sage.

Thomas, J. and Harden, A. (2008) Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8 (1), pp. 45–54.

Thomas, P. (1991) Research models: insiders, gadflies, limestone. In D. S. Anderson and B. J. Biddle (eds) *Knowledge for Policy: Improving Education through Research*. London: Falmer, pp. 225–33.

Thomas, W. I. (1923) *The Unadjusted Girl*. Boston: Little Brown.

Thomas, W. I. (1928) *The Child in America*. New York: Knopf.

Thompson, B. (2001) Significance, effect sizes, stepwise methods, and other issues: strong arguments move the field. *Journal of Experimental Education*, 70 (1), pp. 80–93.

Thompson, B. (2002) What future quantitative social science research could look like: confidence intervals for effect sizes. *Educational Researcher*, 31 (3), pp. 25–32.

Thompson, B. and Snyder, P. A. (1997) Statistical significance testing practices in the Journal of Experimental Education. *Journal of Experimental Education*, 66, pp. 75–83.

Thompson, E. P. (1963) *The Making of the English Working-Class*. London: Victor Gollancz.

Thompson, P. (1978) *The Voice of the Past.* Oxford: Oxford University Press.

Thomson, R. and Holland, J. (2003) Hindsight, foresight and insight: the challenges of longitudinal qualitative research. *International Journal of Social Research Methodology*, 6 (3), pp. 233–44.

Thornberg, R. (2012a) Grounded theory. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 85–93.

Thornberg, R. (2012b) Informed grounded theory. *Scandinavian Journal of Educational Research*, 56 (3), pp. 243–59.

Thorne, B. (1994) *Gender Play: Girls and Boys in School*. New Brunswick, NJ: Rutgers University Press.

Thurstone, L. L. and Chave, E. J. (1929) *The Measurement of Attitudes*. Chicago, IL: University of Chicago Press.

Ticehurst, G. W. and Veal, A. J. (2000) *Business Research Methods*. Frechs Forest, NSW: Pearson.

Tight, M. (2010) The curious case of case study: a viewpoint. *International Journal of Social Research Methodology*, 13 (4), pp. 329–39.

Tillman, L. C. (2002) Culturally sensitive research approaches: an African-American perspective. *Educational Researcher*, 31 (9), pp. 3–12.

Toepoel, V., Vis, C., Das, M. and Van Soest, A. (2009) Design of web questionnaires. *Sociological Methods and Research*, 37 (3), pp. 371–92.

Tombari, M. and Borich, G. (1999) *Authentic Assessment in the Classroom*. Upper Saddle River, NJ: Prentice-Hall.

Torgerson, C. J. (2009) Randomised controlled trials in education research: a case study of an individually randomized pragmatic trial. *Education 3–13*, 37 (4), pp. 313–21.

Torgerson, C. J. and Torgerson, D. J. (2001) The need for randomised controlled trials in educational research. *British Journal of Educational Studies*, 49 (3), pp. 316–28.

Torgerson, C. J. and Torgerson, D. J. (2003a) The design and conduct of randomized controlled trials in education: lessons from health care. *Oxford Review of Education*, 29 (1), pp. 67–80.

Torgerson, D. J. and Torgerson, C. J. (2003b) Avoiding bias in randomized controlled trials in educational research. *British Journal of Educational Studies*, 51 (1), pp. 36–45.

Torgerson, C. J. and Torgerson, D. J. (2008) *Designing Randomised Trials in Health, Education and the Social Sciences*. Houndmills, UK: Palgrave Macmillan.

Torgerson, C. J. and Torgerson, D. J. (2013) *Randomised Trials in Education: An Introductory Handbook*. London: Educational Endowment Foundation. Available from: http://educationendowmentfoundation.org.uk/uploads/pdf/

Randomised_trials_in_education_revised.pdf [Accessed 20 June 2016].
Torrance, H. (2004) Using action research to generate knowledge about educational practice. In G. Thomas and R. Pring (eds) *Evidence-Based Practice in Education*. Buckingham, UK: Open University Press, pp. 187–200.
Torrance, H. (2012) Triangulation, respondent validation and democratic participation in mixed methods research. *Journal of Mixed Methods Research*, 6 (2), pp. 111–23.
Torre, D. and Murphy, J. (2015) A different lens: using photo-elicitation interviews in education research. *Educational Policy Analysis Archives*, 23 (111), pp. 1–26.
Torres, C. A. (1992) Participatory action research and popular education in Latin America. *International Journal of Qualitative Studies in Education*, 5 (1), pp. 51–62.
Tosh, J. (2008) *Why History Matters.* Basingstoke, UK: Palgrave Macmillan.
Tracey, L., Madden, N. A. and Slavin, R. E. (2010) Effects of co-operative learning on the mathematics achievement of years 4 and 5 pupils in Britain: a randomized trial. *Effective Education*, 2 (1), pp. 85–97.
Tracy, S. J. (2010) Qualitative quality: eight 'big-tent' criteria for excellent qualitative research. *Qualitative Inquiry*, 16 (10), pp. 837–51.
Triandis, H. C. (1994) *Culture and Social Behaviour*. New York: McGraw-Hill.
Trifonas, P. P. (2009) Deconstructing research: paradigms lost. *International Journal of Research and Method in Education*, 32 (3), pp. 297–308.
Tripp, D. H. (1993) *Critical Incidents in Teaching.* London: Routledge.
Tripp, D. H. (1994) Teachers' lives, critical incidents and professional practice. *International Journal of Qualitative Studies in Education*, 7 (1), pp. 65–72.
Tripp, D. H. (2003) Action inquiry. *Action Research e-Reports*. Available from: www2.fhs.usyd.edu.au/arow/arer/017.htm#Distinguishing%20action%20research [Accessed 16 April 2010].
Trochim, W. (2006) *Pattern Matching for Construct Validity*. Web Center for Social Research Methods. Available from: www.socialresearchmethods.net/kb/pmconval.php [Accessed 8 December 2013].
Tuckman, B. W. (1972) *Conducting Educational Research*. New York: Harcourt Brace Jovanovich.
Tukey, J. (1962) The future of data analysis. *Annals of Mathematical Statistics*, 33 (1), pp. 1–67.
Tunnicliffe, S. D. and Reiss, M. J. (1999) Talking about Brine Shrimps: three ways of analysing pupil conversations. *Research in Science and Technological Education*, 17 (2), pp. 203–17.
Turkle, S. (2000) Cyborg babies and cy-dough-plasm: ideas about self and life in the culture of simulation. In D. Bell and B. M. Kennedy (eds) *The Cybercultures Reader*. London: Routledge, pp. 547–56.
Turkle, S. (2007) Introduction: the things that matter. In S. Turkle (ed.) *Evocative Objects: Things We Think with*. Cambridge, MA: MIT Press.
Turkle, S. (2015) *Reclaiming Conversation: The Power of Talk in a Digital Age.* New York: Penguin Press.
Turnbull, C. M. (1972) *The Mountain People*. New York: Simon & Schuster Inc.
Tweddle, S., Avis, P., Wright, J. and Waller, T. (1998) Towards evaluating web sites. *British Journal of Educational Technology*, 29 (3), pp. 267–70.
Tymms, P. (1996) Theories, models and simulations: school effectiveness at an impasse. In J. Gray, D. Reynolds, C. T. Fitz-Gibbon and D. Jesson (eds) *Merging Traditions: The Future of Research on School Effectiveness and School Improvement*. London: Cassell, pp. 121–35.
Tymms, P. (2012) Interventions: experiments. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 137–40.
Ulriksen, M. S. and Dadalauri, N. (2016) Single case studies and theory-testing: the knots and dots of the process-tracing method. *International Journal of Social Research Methodology*, 19 (2), pp. 223–39.
Ulysse, B. K. and Lukenchuk, A. (2013) Presaging educational inquiry: historical development of philosophical ideas, traditions, and perspectives. In A. Lukenchuk (ed.) *Paradigms of Research for the 21st Century*. New York: Peter Lang, pp. 3–30.
UNICEF (1989) *United Nations Convention on the Rights of the Child*. London: UNICEF.
United States Department of Education (2002) *Legal and Ethical issues in the Use of V in Education Research*. Working paper 2002–1. Washington, DC: United States Department of Education, Office of Educational Research and Improvement.
University of Loughborough (2009) *Doing a Literature Review*. Available from: http://info.lboro.ac.uk/library/skills/Advice/Litreview.pdf [Accessed 6 February 2010].
University of North Carolina (2007) *Literature Reviews*. Available from: www.unc.edu/depts/wcweb/handouts/literature_review.html [Accessed 6 February 2010].
Uprichard, E. (2012) Dirty data: longitudinal classification systems. *The Sociological Review*, 59 (2), pp. 93–112.
Uprichard, E. (2013) Sampling: bridging probability and non-probability designs. *International Journal of Asocial Research Methodology*, 16 (1), pp. 1–11.
US Dept of Health, Education and Welfare, Public Health Service and National Institutes of Health (1971) *The Institutional Guide to D.H.E.W. Policy on Protecting Human Subjects*. DHEW Publication (NIH), 2 December, pp. 72–102.
Usher, P. (1996) Feminist approaches to research. In D. Scott and R. Usher (eds) *Understanding Educational Research*. London: Routledge, pp. 120–42.
Valentine, J. C. (2009) Judging the quality of primary research. In H. M. Cooper, L. V. Hedges and J. C. Valentine (eds) *The Handbook of Research Synthesis and Meta-analysis* (second edition). New York: The Russell Sage Foundation, pp. 129–46.
Vallerand, R. J. (1989) Vers une méthodologie de validation trans-culturelle de questionnaires psychologiques (Toward a methodology of cross-cultural validation of psychological questionnaires). *Psychologie Canadienne*, 30 (4), pp. 662–80.
Vallerand, R. J., Pelletier, L. G., Blais, M. R., Brière, N. M., Senécal, C. and Vallières, E. F. (1992) The academic motivation scale: a measure of intrinsic, extrinsic and amotivation in education. *Educational and Psychological Measurement*, 52 (4), pp. 1003–17.

Van den Hoven, M. J. (1997) Privacy and the varieties of moral wrong-doing in an information age. *Computers and Society*, 27 (3), pp. 33–7.

Van Maanen, J. (1988) *Tales of the Field: On Writing Ethnography*. Chicago, IL: University of Chicago Press.

Van Meter, K. M. (2000) Sensitive topics – sensitive questions: overview of the sociological research literature. *Bulletin de Méthodologie Sociologique*, 68 (1), pp. 59–79.

van Rekom, J. and Wierenga, B. (2007) On the hierarchical nature of means–end relationships in laddering data. *Journal of Business Research*, 60 (4), pp. 401–10.

Vartanian, T. P. (2011) *Secondary Data Analysis*. New York: Oxford University Press.

Vaughan, E. (2012) Discourse analysis. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 272–82.

Verma, G. K. and Mallick, K. (1999) *Researching Education: Perspectives and Techniques*. London: Falmer.

Vermunt, J. K. and Magidson, J. (2002) Latent Class Cluster Analysis. In J. A. Hagenaars and A. L. McCutcheon (eds) *Applied Latent Class Analysis*. Cambridge: Cambridge University Press, pp. 89–106.

Verschuren, P. J. M. (2003) Case study as a research strategy: some ambiguities and opportunities. *International Journal of Research Methodology*, 6 (2), pp. 121–39.

Vignoles, A. (2007) *The Use of Large Scale Data-Sets in Educational Research*. London: Teaching and Learning Research Programme. Available from: www.tlrp.org/capacity/rm/wt/vignoles [Accessed 6 March 2016].

Voss, R., Thorsten, G. and Szmigin, I. (2007) Service quality in higher education: the role of student expectations. *Journal of Business Research*, 60 (9), pp. 949–59.

Vulliamy, G. (1990) The potential of qualitative educational research in developing countries. In G. Vulliamy, K. Lewin and D. Stephens (1990) *Doing Educational Research in Developing Countries: Qualitative Strategies*. London: Falmer, pp. 7–25.

Vulliamy, G., Lewin, K. and Stephens, D. (1990) *Doing Educational Research in Developing Countries: Qualitative Strategies*. London: Falmer.

Wadensjö, C. (2014) Perspectives on role play: analysis, training and assessments. *The Interpreter and Translator Trainer*, 8 (3), pp. 437–51.

Wagner, B. J. (1998) Drama as a way of knowing. In J. Saxton and C. Miller (eds) *The Research of Practice: The Practice of Research*. Victoria, BC: International Drama in Education Research Institute, pp. 55–72.

Wainer, H. (ed.) (1990) *Computerized Adaptive Testing: A Primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H. (2015) *Computerized Adaptive Testing* (second edition). New York: Routledge.

Wainer, H. and Dorans, N. J. (2000) *Computerized Adaptive Testing: A Primer* (second edition). Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H. and Mislevy, R. J. (1990) Item response theory, item calibration and proficiency estimation. In H. Wainer (ed.) *Computerized Adaptive Testing: A Primer*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 65–102.

Waldrop, M. M. (1992) *Complexity: The Emerging Science at the Edge of Order and Chaos*. Harmondsworth: Penguin.

Walford, G. (1994a) A new focus on the powerful. In G. Walford (ed.) *Researching the Powerful in Education*. London: UCL Press, pp. 2–11.

Walford, G. (1994b) Ethics and power in a study of pressure group politics. In G. Walford (ed.) *Researching the Powerful in Education*. London: UCL Press, pp. 81–93.

Walford, G. (1994c) Reflections on researching the powerful. In G. Walford (ed.) *Researching the Powerful in Education*. London: UCL Press, pp. 222–31.

Walford, G. (2001) *Doing Qualitative Educational Research: A Personal Guide to the Research Process*. London: Continuum.

Walford, G. (2005) Research ethical guidelines and anonymity. *International Journal of Research and Method in Education*, 28 (1), pp. 83–93.

Walford, G. (2009) For ethnography. *Ethnography and Education*, 4 (3), pp. 271–82.

Walford, G. (2012) Researching the powerful in education: a re-assessment of the problems. *International Journal for Research and Method in Education*, 35 (2), pp. 111–18.

Walker, R. (1980) Making sense and losing meaning: problems of selection in doing Case Study. In H. Simons (ed.) *Towards a Science of the Singular*. Norwich, UK: University of East Anglia, Centre for Applied Research in Education, pp. 222–35.

Wall, C. (2008) Picturing an occupational identity: images of teachers in careers and trade union publications 1940–2000. *History of Education*, 37 (2), pp. 317–40.

Wall, K., Hall, E. and Woolner, P. (2012) Visual methodology: previously, now and in the future. *International Journal of Research and Method in Education*, 35 (3), pp. 223–6.

Wall, S. (2006) An autoethnography on learning about autoethnography. *International Journal of Qualitative Methods*, 5 (2), pp. 1–12.

Wall, S. (2008) Easier said than done: writing an autoethnography. *International Journal of Qualitative Methods*, 7 (1), pp. 38–53.

Waller, D., Hunt, E. and Knapp, D. (1998) The transfer of spatial knowledge in virtual environment training. *Presence*, 7 (2), pp. 129–43.

Walsh, S. (2006) *Investigating Classroom Discourse*. London: Routledge.

Waltz, M. (2007) The relationship of ethics to quality: a particular case of research in autism. *International Journal of Research and Method in Education*, 30 (3), pp. 353–61.

Wang, J. (2008) Effect size and practical importance: a non-monotonic match. *International Journal of Research and Method in Education*, 31 (2), pp. 125–32.

Wardekker, W. L. and Miedama, S. (1997) Critical pedagogy: an evaluation and a direction for reformulation. *Curriculum Inquiry*, 27 (1), pp. 45–61.

Waring, M. (2012) Grounded theory. In J. Arthur, M. Waring, R. Coe and L. V. Hedges (eds) *Research Methods and Methodologies in Education*. London: Sage, pp. 297–308.

Warschauer, M. and Matuchniak, T. (2010) New technology and digital worlds: analysing evidence of equity in access, use, and outcomes. *Review of Research in Education*, 34 (1), pp. 179–225.

Waterman, A. H., Blades, M. and Spencer, C. (2001) Interviewing children and adults: the effect of question format

on the tendency to speculate. *Applied Cognitive Psychology*, 15 (5), pp. 521–31.

Waters, B. (2016) 'A part to play': the value of role-play simulation in undergraduate legal education. *The Law Teacher*, 50 (2), pp. 172–94.

Watkins, D. A. (2007) Comparing ways of learning. In M. Bray, R. Adamson and M. Mason (eds) *Comparative Education Research: Approach and Methods*. Hong Kong: Comparative Education Research Centre, University of Hong Kong, pp. 299–313.

Watson, M., Jones, D. and Burns, L. (2007) Internet research and informed consent: an ethical model for using archived e-mails. *International Journal of Therapy and Rehabilitation*, 14 (9), pp. 396–403.

Watt, J. H. (1997) Using the Internet for quantitative survey research. *Quirk's Marketing Research Review*, July. Available from: www.swiftinteractive.com.white1.asp [Accessed 6 January 2003].

Watts, H. (1985) When teachers are researchers, teachers improve. *Journal of Staff Development*, 6 (2), pp. 118–27.

Watts, J. H. (2011) Ethical and practical challenges of participant observation in sensitive health research. *International Journal of Social Research Methodology*, 14 (4), pp. 301–12.

Watts, M. (2007) They have tied me to a stake: reflections on the art of case study research. *Qualitative Inquiry*, 13 (2), pp. 204–17.

Watts, M. and Ebbutt, D. (1987) More than the sum of the parts: research methods in group interviewing. *British Educational Research Journal*, 13 (1), pp. 25–34.

Wax, M. (1982) Research reciprocity rather than informed consent in fieldwork. In J. Sieber (ed.) *The Ethics of Social Research: Fieldwork, Regulation and Publication*. New York: Springer-Verlag, pp. 33–48.

Webb, G. (1996) Becoming critical of action research for development. In O. Zuber-Skerritt (ed.) *New Directions in Action Research*. London: Falmer, pp. 137–61.

Webb, L. M., Walker, K. L. and Bollis, T. S. (2004) Feminist pedagogy in the teaching of social research methods. *International Journal of Social Research Methodology*, 7 (5), pp. 415–28.

Weber, M. (1949) Objectivity in social science and social policy. In M. Weber (ed.) *The Methodology of the Social Sciences* (trans. E. A. Shils and H. A. Finch). New York: Free Press, pp. 49–112.

Weber, R. P. (1990) *Basic Content Analysis* (second edition). Thousand Oaks, CA: Sage.

Weber, S. and Mitchell, C. (1995) *'That's Funny, You Don't Look Like a Teacher': Interrogating Images and Identity in Popular Culture*. London: RoutledgeFalmer.

Webster, J. P. and da Silva, S. M. (2013) Doing educational ethnography in an online world: methodological challenges, choices and innovations. *Ethnography and Education*, 8 (2), pp. 123–30.

Webster, R. (2015) The classroom experiences of pupils with special educational needs in mainstream primary schools: 1976 to 2012 – what do data from systematic observation studies reveal about pupils' educational experiences over time? *British Educational Research Journal*, 41 (6), pp. 992–1009.

Wedeen, P., Winter, J. and Broadfoot, P. (2002) *Assessment: What's in It for Schools?* London: RoutledgeFalmer.

Weems, G. H., Onwuegbuzie, A. J. and Lustig, D. (2003) Profiles of respondents who respond inconsistently to positively- and negatively-worded items on rating scales. *Evaluation and Research in Education*, 17 (1), pp. 45–60.

Weiler, K. (1998) *Country Schoolwomen: Teaching in Rural California 1850–1950*. Stanford, CA: Stanford University Press.

Weisberg, H. F., Krosnick, J. A. and Bowen, B. D. (1996) *An Introduction to Survey Research, Polling, and Data Analysis* (third edition). Thousand Oaks, CA: Sage.

Weiskopf, R. and Laske, S. (1996) Emancipatory action research: a critical alternative to personnel development or a new way of patronising people? In O. Zuber-Skerritt (ed.) *New Directions in Action Research*. London: Falmer, pp. 121–36.

Weiss, C. (1991a) The many meanings of research utilization. In D. S. Anderson and B. J. Biddle (eds) *Knowledge for Policy: Improving Education through Research*. London: Falmer, pp. 173–82.

Weiss, C. (1991b) Knowledge creep and decision accretion. In D. S. Anderson and B. J. Biddle (eds) *Knowledge for Policy: Improving Education through Research*. London: Falmer, pp. 183–92.

Wellington, J. (2015) *Educational Research* (second edition). London: Bloomsbury Academic.

Wetherell, M., Taylor, S. and Yates, S. (2001) *Discourse as Data*. London: Sage.

Wheatley, M. (1999) *Leadership and the New Science: Discovering Order in a Chaotic World* (second edition). San Francisco, CA: Berrett-Koehler Publishers.

White, P. (2009) *Developing Research Questions: A Guide for Social Scientists*. London: Palgrave.

White, P. (2013) Who's afraid of research questions? The neglect of research questions in the methods literature and a call for question-led methods teaching. *International Journal of Research and Method in Education*, 36 (3), pp. 213–27.

Whitehead, J. (1985) An analysis of an individual's educational development: the basis for personally oriented action research. In M. Shipman (ed.) *Educational Research: Principles, Policies and Practices*. Lewes, UK: Falmer, pp. 97–108.

Whitty, G. and Edwards, A. D. (1994) Researching Thatcherite policy. In G. Walford (ed.) *Researching the Powerful in Education*. London: UCL Press, pp. 14–31.

Whyte, W. F. (1955) *Street Corner Society: The Social Structure of an Italian Slum* (second edition). Chicago, IL: University of Chicago Press.

Whyte, W. F. (1982) Interviewing in field research. In R. Burgess (ed.) *Field Research: A Sourcebook and Field Manual*. London: Allen & Unwin, pp. 111–22.

Whyte, W. F. (1993) *Street Corner Society: The Social Structure of an Italian Slum* (fourth edition). Chicago, IL: University of Chicago Press.

Wickens, P. (1987) *The Road to Nissan: Flexibility, Quality, Teamwork*. Basingstoke, UK: Macmillan.

Wideen, M., Mayer-Smith, J. and Moon, B. (1998) A critical analysis of the research on learning to teach: making the case for an ecological perspective on inquiry. *Review of Educational Research*, 68 (2), pp. 130–78.

Wiggins, G. (1998) *Educative Assessment*. San Francisco, CA: Jossey-Bass.

Wilde, O. (2008) *Lord Arthur Savile's Crime and Other Stories*. Whitefish, MT: Kessinger Publishing [1891].

Wiles, J. and Bondi, J. C. (1984) *Curriculum Development: A Guide to Practice* (second edition). Columbus, OH: Charles E. Merrill Publishing.

Wiles, J. C. and Bondi, J. C. (2014) *Curriculum Development: A Guide to Practice* (ninth edition). New York: Pearson.

Wiles, R., Crow, G., Heath, S. and Charles, V. (2008a) The management of confidentiality and anonymity in social research. *International Journal of Social Research Methodology*, 11 (5), pp. 417–28.

Wiles, R., Prosser, J., Bagnoli, A., Clark, A., Davies, K., Holland, S. and Renold, E. (2008b) *Visual Ethics: Ethical Issues in Visual Research*. Working Paper NCRM/011 for ESRC National Centre for Research Methods. Manchester: ESRC National Centre for Research Methods. Available from: http://eprints.ncrm.ac.uk/421/1/MethodsReviewPaperNCRM-011.pdf [Accessed 18 May 2010].

Wilkinson, J. (2000) Direct observation. In G. M. Breakwell, S. Hammond and C. Fife-Shaw (eds) *Research Methods in Psychology* (second edition). London: Sage, pp. 224–38.

Wilkinson, L. and the Task Force on Statistical Inference, APA Board of Scientific Affairs (1999) Statistical methods in psychology journals: guidelines and explanations. *American Psychologist*, 54 (8), pp. 594–604.

Williams, D., Brick, J. M., Montaquila, J. M. and Han, D. F. (2016) Effects of screening questionnaires on response in a two-phase postal survey. *International Journal of Social Research Methodology*, 19 (1), pp. 51–67.

Williams, R. (1961) *The Long Revolution*. Harmondsworth, UK: Penguin.

Williams, S. G. (2012) The ethics of Internet research. *Online Journal of Nursing Informatics (OJNI)*, 16 (2). Available from: http://ojni.org/issues/?p=1708 [Accessed 8 April 2016].

Willis, J. and Saunders, M. (2007) Research in a post-colonial world: the example of Australian Aborigines. In M. Pitt and A. Smith (eds) *Researching the Margins*. London: Palgrave Macmillan, pp. 96–114.

Willis, P. E. (1977) *Learning to Labour*. Farnborough, UK: Saxon House.

Wilson, C. and Powell, M. (2001) *A Guide to Interviewing Children: Essential Skills for Counsellors, Police Lawyers and Social Workers*. London: Routledge.

Wilson, D. B. (2009) Systematic coding. In H. M. Cooper, L. V. Hedges and J. C. Valentine (eds) *The Handbook of Research Synthesis and Meta-analysis* (second edition). New York: The Russell Sage Foundation, pp. 159–76.

Wilson, I., Huttly, S. R. A. and Fenn, B. (2006) A case study of sample design for longitudinal research: Young Lives. *International Journal of Social Research Methodology*, 9 (5), pp. 351–65.

Wilson, M. (1996) Asking questions. In R. Sapsford and V. Jupp (eds) *Data Collection and Analysis*. London: Sage and the Open University Press, pp. 94–120.

Wilson, N. and McLean, S. (1994) *Questionnaire Design: A Practical Introduction*. Newtown Abbey, Co. Antrim: University of Ulster Press.

Windle, P. E. (2010) Secondary data analysis: is it useful and valid? *Journal of PeriAnesthesia Nursing*, 25 (5), pp. 322–4.

Windschitl, M. (2002) Framing constructivism in practice as the negotiation of dilemmas: an analysis of the conceptual, pedagogical, cultural, and political challenges facing teachers. *Review of Educational Research*, 72 (2), pp. 131–75.

Wineburg, S. S. (1991) The self-fulfilment of the self-fulfilling prophecy. In D. S. Anderson and B. J. Biddle (eds) *Knowledge for Policy: Improving Education through Research*. London: Falmer, pp. 276–90.

Winter, D. A., Bell, R. C. and Watson, S. (2010) Midpoint ratings on personal constructs: constriction or the middle way? *Journal of Constructivist Psychology*, 23 (4), pp. 337–56.

Winter, G. (2000) A comparative discussion of the notion of 'validity' in qualitative and quantitative research. *The Qualitative Report*, 4 (3–4), March. Available from: www.nova.edu/sss/QR/QR4-3/winter.html [Accessed 29 October 2005].

Winter, R. (1996) Some principles and procedures for the conduct of action research. In O. Zuber-Skerritt (ed.) *New Directions in Action Research*. London: Falmer, pp. 13–27.

Witmer, D. F., Colman, R. W. and Katzman, S. L. (1999) From paper-and-pencil to screen-and-keyboard: toward a methodology for survey research on the Internet. In S. Jones (ed.) *Doing Internet Research*. Thousand Oaks, CA: Sage, pp. 145–61.

Wittgenstein, L. (1974) *Tractatus Logico-Philosophicus* (trans. D. Pears and B. McGuiness). London: Routledge & Kegan Paul.

Witzel, A. (2000) The problem-centered interview. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research [Online Journal]*, 1 (1), pp. 1–9, article 22. Available from: www.qualitative-research.net/index.php/fqs/article/view/1132/2522 [Accessed 6 March 2003].

Wolcott, H. F. (1973) *The Man in the Principal's Office*. New York: Holt, Rinehart & Winston.

Wolcott, H. F. (1992) Posturing in qualitative research. In M. LeCompte, W. L. Millroy and J. Preissle (eds) *The Handbook of Qualitative Research in Education*. London: Academic Press Ltd, pp. 3–52.

Wolcott, H. F. (1994) *Transforming Qualitative Data: Description, Analysis and Interpretation*. Thousand Oaks, CA: Sage.

Wolf, R. M. (1994) The validity and reliability of outcome measure. In A. C. Tuijnman and T. N. Postlethwaite (eds) *Monitoring the Standards of Education*. Oxford: Pergamon, pp. 121–32.

Wolff, S. (2004) Ways into the field and their variants. In U. Flick, E. von Kardoff and I. Steinke (eds) *A Companion to Qualitative Research*. London: Sage, pp. 195–202.

Wood, S. (1980) Reactions to redundancy. Unpublished PhD thesis, University of Manchester. Quoted in R. M. Lee (ed.) (1993) *Doing Research on Sensitive Topics*. London: Sage.

Woods, D. (2010) *Transana Keyboard Shortcuts and Transcript Notation* (originally by R. Henne and subsequently modified by D. Woods). Available from: www.transana.org/images/TransanaShortcuts.pdf [Accessed 1 May 2010].

Woods, M., Macklin, R. and Lewis, G. K. (2016) Researcher reflexivity: exploring the impacts of CAQDAS use. *International Journal of Social Research Methodology*, 19 (4), pp. 385–403.

Woods, P. (1979) *The Divided School*. London: Routledge & Kegan Paul.
Woods, P. (1983) *Sociology and the School*. London: Routledge & Kegan Paul.
Woods, P. (1986) *Inside Schools: Ethnography in Educational Research*. London: Routledge & Kegan Paul.
Woods, P. (1989) *Working for Teacher Development*. Dereham, UK: Peter Francis Publishers.
Woods, P. (1992) Symbolic interactionism: theory and method. In M. LeCompte, W. L. Millroy and J. Preissle (eds) *The Handbook of Qualitative Research in Education*. London: Academic Press Ltd, pp. 337–404.
Wragg, E. C. (1994) *An Introduction to Classroom Observation*. London: Routledge.
Wragg, E. C. (2002) Interviewing. In M. Coleman and A. R. J. Briggs (eds) *Research Methods in Educational Leadership*. London: Paul Chapman Publishing, pp. 143–58.
Wright, D. B. (2003) Making friends with your data: improving how statistics are conducted and reported. *British Journal of Educational Psychology*, 73 (1), pp. 123–36.
Wright, R. and Powell, M. B. (2006) Investigative interviewers' perceptions of their difficulty in adhering to open-ended questions with child witnesses. *International Journal of Police Science and Management*, 8 (4), pp. 316–25.
Wright, R. P. and Lam, S. S. K. (2002) Comparing apples with apples: the importance of element wording in grid applications. *Journal of Constructivist Psychology*, 15 (2), pp. 109–19.
Yazan, B. (2015) Three approaches to case study methods in education: Yin, Merriam and Stake. *The Qualitative Report*, 20 (2), pp. 134–52.
Yeung, K. W. and Watkins, D. (2000) Hong Kong student teachers' personal construction of teaching efficacy. *Educational Psychology*, 20 (2), pp. 213–35.
Yim, O. and Ramdeen, K. T. (2015) Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. *The Quantitative Methods for Psychology*, 11 (1), pp. 8–21.
Yin, R. K. (2006) Mixed methods research: are the methods genuinely integrated or merely parallel? *Research in Schools*, 13 (1), pp. 41–7.
Yin, R. K. (2009) *Case Study Research: Design and Methods* (fourth edition). Thousand Oaks, CA: Sage.
Yorke, D. M. (1978) Repertory grids in educational research: some methodological considerations. *British Educational Research Journal*, 4 (2), pp. 63–74.
Yorke, M. (1983) Straight or bent? An inquiry into rating scales in repertory grids. *British Educational Research Journal*, 9 (2), pp. 141–51.
Yorke, M. (2011) Analysing existing datasets: some considerations arising from practical experience. *International Journal of Research and Method in Education*, 34 (3), pp. 255–67.
Youngman, M. B. (1984) Designing questionnaires. In J. Bell, T. Bush, A. Fox, J. Goodey and S. Goulding (eds) *Conducting Small-Scale Investigations in Educational Management*. London: Harper & Row, pp. 156–76.
Yu, K. (2011) Exploring the nature of the researcher–practitioner relationship in qualitative educational research publications. *International Journal of Qualitative Studies in Education*, 24 (7), pp. 785–804.
Zhang, L., Beach, R. and Sheng, Y. (2016) Understanding the use of online role-play for collaborative argument through teacher experiencing: a case study. *Asia-Pacific Journal of Teacher Education*, 44 (3), pp. 242–56.
Zhao, S. (1991) Meta-theory, meta-method, meta-data-analysis: what, why, and how? *Sociological Perspectives*, 34 (3), pp. 377–90.
Zhao, S. (2003) 'Being there' and the role of presence technology. In G. Riva, F. Davide and W. A. Ijsselsteijn (eds) *Being There: Concepts, Effects and Measurements of User Presence in Synthetic Environments*. Amsterdam: International Operations Press, pp. 137–46.
Zhao, Y. (2014) *Who's Afraid of the Big Bad Dragon? Why China Has the Best (and Worst) Education System in the World*. San Francisco, CA: Jossey-Bass.
Ziliak, S. T. and McCloskey, D. N. (2008) *The Cult of Statistical Significance*. Ann Arbor, MI: University of Michigan Press.
Zimbardo, P. G. (1973) On the ethics of intervention in human psychological research with specific reference to the 'Stanford Prison Experiment'. *Cognition*, 2 (2), pp. 243–56.
Zimbardo, P. C. (1984) On the ethics of intervention in human psychological research with specific reference to the 'Stanford Prison Experiment'. In J. Murphy, M. John and H. Brown (eds) *Dialogues and Debates in Social Psychology*. London: Lawrence Erlbaum with the Open University Press.
Zimbardo, P. G. (2007a) *The Lucifer Effect: Understanding How Good People Turn Evil*. New York: Random House.
Zimbardo, P. G. (2007b) Revisiting the Stanford Prison Experiment: a lesson in the power of situation. *Chronicle of Higher Education*, 30 March, pp. B6–B7.
Zimbardo, P. G. (2008) From the Bronx to Stanford to Abu Ghraib. In R. V. Levine, A. Rodriques and L. Zelezny (eds) *Journeys in Social Psychology: Looking Back to Inspire the Future*. New York: Taylor & Francis, pp. 85–104.
Zimbardo, P. G., Maslach, C. and Haney, C. (2000) Reflections on the Stanford Prison Experiment: genesis, transformations, consequences. In T. Blass (ed.) *Obedience to Authority: Current Perspectives on the Milgram Paradigm*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 193–237.
Zimmer, L. (2006) Qualitative meta-synthesis: a question of dialoguing with texts. *Journal of Advanced Nursing*, 53 (3), pp. 311–18.
Zimmer, M. (2010) 'But the data is already public': on the ethics of research in Facebook. *Ethics and Information Technology*, 12 (4), pp. 313–25.
Znaniecki, F. (1934) *The Method of Sociology*. New York: Farrar & Rinehart.
Zuber-Skerritt, O. (1996a) Introduction. In O. Zuber-Skerritt (ed.) *New Directions in Action Research*. London: Falmer, pp. 3–9.
Zuber-Skerritt, O. (1996b) Emancipatory action research for organisational change and management development. In O. Zuber-Skerritt (ed.) *New Directions in Action Research*. London: Falmer, pp. 83–105.

Index