The Magenta Book: guidance notes for policy evaluation and analysis

Government Social Research Unit

HM Treasury
1 Horse Guards Road
London SW1A 2HQ

Contents

The Magenta Book: guidance notes for policy evaluation and analysis

Introduction

Published: July 2003
Updated: October 2007

Government Social Research Unit

HM Treasury
1 Horse Guards Road
London SW1A 2HQ

Introduction

Objectives, format and uses

The Magenta Book is not another textbook on policy evaluation and analysis. The field has plenty of such texts and these will be referred to throughout The Magenta Book. Rather, The Magenta Book is a set of guidance notes for policy evaluators and analysts, and people who use and commission policy evaluation. It has a strong focus on policy evaluation *in government* and is structured to meet the needs of government analysts and policy makers. It is hoped that it may also meet the needs of analysts and users of evaluation outside of government, and that it will stimulate dialogue and collaboration between the worlds of government, academia and the wider research and evaluation community.

To meet the needs of different audiences, each chapter of The Magenta Book consists of a set of *guidance notes* and a *background paper*. *The* guidance notes offer a summary of key issues and enable the reader to quickly access further sources of relevant information while the background papers explore the issues covered in greater depth for those who are interested in a more detailed discussion of methodological issues. The guidance notes for each chapter can be found on the Policy Hub website (http://www.policyhub.gov.uk/). The background papers are available in PDF format and can be downloaded via links at the end of this chapter, from the Government Social Research website (http://www.gsr.gov.uk) and via links from Policy Hub.

The Magenta Book has been developed in the context of the demands of evidence-based policy making and the changing needs of analysis in and for government. A series of publications since 1997, including the Modernising Government White Paper (Cabinet Office 1999a), Policy Making for the 21st Century (Cabinet Office, 1999b, Adding-it-Up (Cabinet Office, 2000), and Better Policy Making (Cabinet Office, 2001) have stressed the importance of sound evidence, proper evaluation and good analysis at the heart of policy making. This, in turn, has generated a demand for guidance on how to undertake high quality evaluation, appraisal and analysis for policy making. This demand has been met by a number of important documents including a revised version of H.M Treasury's Evaluation and Appraisal for Government (The Green Book). The Better Regulation Executive has developed an Impact Assessment tool which can help policy makers think through the consequences of Government interventions in the public, private and third sectors and enable Government to weigh and present the relevant evidence on the positive and negative effects of such interventions.

The Magenta Book complements these other sources of guidance by providing a user-friendly guide for specialists and generalists alike on the

methods used by social researchers when they commission, undertake and manage policy research and evaluation. The Magenta Book endeavours to provide guidance on social research methods for policy evaluation in readable and understandable language. Where technical detail is required, or it is necessary to expand on methodological procedures and arguments, these are presented in boxed and shaded areas. The Magenta Book provides examples of evaluations that have used the available methods appropriately and effectively, and it highlights what it is that is good about them.

The Magenta Book covers the broad range of methods used in policy evaluation, and the approaches of different academic disciplines (social policy, sociology, economics, statistics, operational research). The Magenta Book is driven by the substantive policy questions being asked of analysts, rather than by methodological disputes between academic disciplines or different schools of thought. The Magenta Book includes guidance on how to use summative and formative, quantitative and qualitative, experimental and experiential methods of policy evaluation appropriately and effectively.

The Magenta Book is organised around a number of questions that are frequently asked about policy evaluation and analysis (see below). In answering these questions *The Magenta Book* provides guidance on:

- How to refine a policy question to get a useful answer
- The main evaluation methods that are used to answer policy questions
- The strengths and weaknesses of different methods of evaluation
- The difficulties that arise in using different methods of evaluation
- The benefits that are to be gained from using different methods of evaluation
- Where to go to find out more detailed information about policy evaluation and analysis

The Magenta Book is published electronically and in installments until the complete set of evaluation questions that it addresses has been covered. Electronic publication will enable fast production and dissemination, and its contents to be updated regularly. Electronic publication also allows some degree of interactive use in that readers can respond to its contents, challenge its guidance, raise new questions, offer new insights, and contribute to its regular updating.

The Magenta Book is linked to a series of training and professional development modules in Policy Evaluation and Analysis that have been developed by the Government Social Research Unit (GSRU) for

government analysts and policy makers who use and commission policy evaluation http://www.hm-treasury.gsi.gov.uk/gsr/gsru_courses/gsru_courses_2006-7.asp (GSI only). These modules are built around the principles of problem-based learning and problem-based assessment, so that analysts can develop their analytical skills by answering 'real life' analytical and policy problems that arise in their everyday work. GSRU have also developed a Masters qualification in Policy Analysis and Evaluation which is run jointly with the Institute of Education, University of London. For details see: http://www.gsr.gov.uk/professional_development/msc/index.asp.

The Magenta Book: guidance notes for policy evaluation and analysis

Background paper 1: what is policy evaluation?

Published July 2003
Updated October 2007

Government Social Research Unit

HM Treasury
1 Horse Guards Road
London SW1A 2HQ

1 What is policy evaluation?

1.1 The demand for policy evaluation

The need for good analysis and sound evaluation to be at the heart of policy making has been recognised in a number of government publications (Cabinet Office, 1999a, 1999b, 2000, 2001). The Adding-It-Up Report (Cabinet Office 2000), for instance, argued that:

> *"Rigorous analysis and, where appropriate, modelling is in the best interests of both Ministers and senior officials. They lead to better decisions and improved policy outcomes. Without soundly based analysis and modelling, those involved in the formulation of policy and the delivery of services will work in the dark. As a result, the pace of reform may be slow."*
>
> (Cabinet Office 2000: 3)

Some guidance on evaluation and appraisal is already available from within Government. HM Treasury produces a guide on economic appraisal and analysis, known as The Green Book.. This distinguishes between *ex ante* appraisal of policy options and *ex post* evaluation of policies that have been implemented. The Green Book is mostly concerned with evaluating policies, programmes and projects using economic appraisal techniques. Policy evaluation across government, however, has a wider meaning and uses a variety of analytical tools and methodological procedures from a wide range of academic disciplines. This is the focus of The Magenta Book.

Other guidance produced by Government includes the Better Regulation Executive's Impact Assessment tool which can help policy makers think through the consequences of Government interventions in the public, private and third sectors and enable Government to weigh and present the relevant evidence on the positive and negative effects of such interventions, and Assessing the Impacts of Spatial Interventions produced by ODPM to provide guidance "on the assessment of interventions with a spatial focus (typically regeneration, renewal or regional development initiatives)" (ODPM, 2004:5).

The Magenta Book complements all of these sources by providing guidance for social researchers, other analysts, and policy makers on the wide range of evaluation methods used in policy evaluation.

1.2 What is evaluation?

Evaluation has been defined as a family of research methods which seeks "to systematically investigate the effectiveness of social interventions….in ways that improve social conditions" (Rossi, Freeman and Lipsey, 1999:20). Another definition of evaluation is "the process of determining the merit, worth, or value of something, or the product of

that process" (Scriven, 1991). Drawing upon these two sources the following definition of policy evaluation can be proposed:

> *"Policy evaluation uses a range of research methods to systematically investigate the effectiveness of policy interventions, implementation and processes, and to determine their merit, worth, or value in terms of improving the social and economic conditions of different stakeholders."*

The importance of a range of research methods is paramount. Policy evaluation uses quantitative and qualitative methods, experimental and non-experimental designs, descriptive and experiential methods, theory based approaches, research synthesis methods, and economic evaluation methods. It privileges no single method of inquiry and acknowledges the complementary potential of different research methods. The methods used in policy evaluation and analysis are usually driven by the substantive issues at hand rather than *a priori* preferences (Greene, Benjamin and Goodyear, 2001).

1.3 What types of evaluation are used in government?

1.3.1 Summative and formative evaluation

Two types of evaluation that are commonly used in government are *summative* and *formative* evaluation. *Summative* evaluation, which is sometimes referred to as impact evaluation, asks questions such as: What impact, if any, does a policy, programme or some other type of government intervention have in terms of specific outcomes for different groups of people? It seeks to provide estimates of the effects of a policy either in terms of what was expected of it at the outset, or compared with some other intervention, or with doing nothing at all (i.e. the counterfactual).

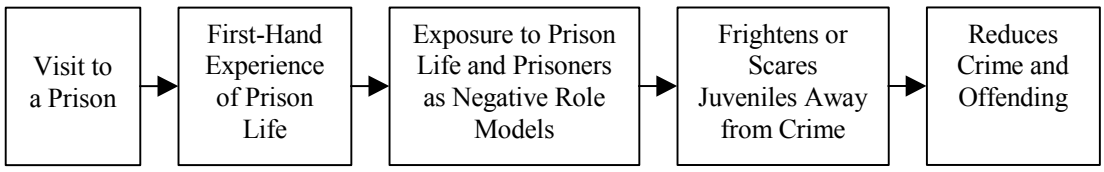
Formative evaluation, which is sometimes referred to as *process* evaluation, asks *how, why,* and *under what conditions* does a policy intervention (or a programme, or a project) work, or fail to work? These questions are important in determining the effective development (i.e. formation), implementation and delivery of policies, programmes or projects. Formative evaluation typically seeks information on the *contextual* factors, mechanisms and processes underlying a policy's success or failure. This often involves addressing questions such as *for whom* a policy has worked (or not worked), and why.

This distinction between summative and formative evaluations is not always as rigid as the above characterisation might suggest. Proponents of the *Theories of Change* approach to evaluation (Chen, 1990; Connell et al, 1995; Funnel, 1997, Owen and Rodgers, 1999; Weiss, 1997; Judge and Bauld, 2001) would argue that determining whether or not a policy has worked, or has been effective, necessarily involves asking questions about *how* it has worked, for *whom, why,* and *under what conditions* it has

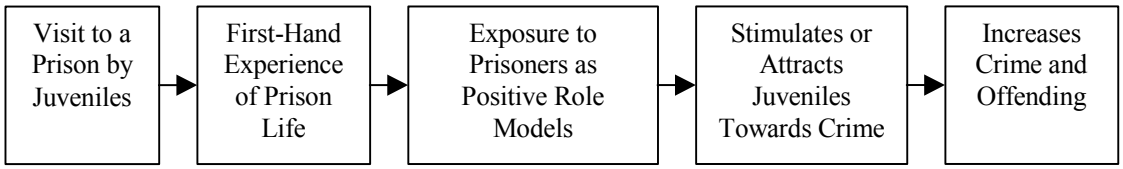
worked or not worked. Nonetheless, the contrast between evaluating *whether* a policy intervention has been effective (summative evaluation), and *why* it has done so (formative evaluation), is one that is conventionally made in the policy evaluation literature.

1.3.2 Theory-based evaluation

Theory-Based approaches to evaluation, which include the *Theories of Change* approach mentioned above, as well as *Programme Theory Evaluation* (Rogers *et al*, 2000) and some aspects of *Realistic Evaluation* (Pawson and Tilley, 1997), focus on unpacking the theoretical or logical sequence by which a policy intervention is expected to bring about its desired effects. Theory-Based approaches attempt to identify the *mechanisms* by which policies and/or programmes might produce their effects. For instance, the common underlying theory of the juvenile awareness programmes for preventing juvenile delinquency (such as the 'Scared Straight' programmes in the United States, (Petrosino, Turpin-Petrossino, and Buehler, 2002)) suggest the following sequential steps:



An alternative possible sequence of outcomes, which can be tested empirically, might be as follows:



Failure to be clear about the causal sequence by which a policy is expected to work can result in well intentioned policies being misplaced, and outcomes that are contrary to those that were anticipated. Theory-Based evaluation provides a number of ways of carrying out an analysis of the logical or theoretical consequences of a policy, and can increase the likelihood of the desired outcome being achieved. Theory-Based initiatives will be considered in greater detail in a forthcoming section of the Magenta Book.

1.3.3 Can the policy, programme or project be evaluated?

Another important question to ask is whether or not a policy, programme or project can be evaluated at all. Some policy initiatives

and programmes can be so complicated and diffuse that they have little prospect of meeting the central requirements of evaluability. These are that the interventions, and the target population, are clear and identifiable; that the outcomes are clear, specific and measurable; and that an appropriate evaluation design can be implemented (Patton, 1990).

1.3.4 Have the goals of a policy, programme or project been achieved? – goals-based evaluation

This is one of the most frequently asked questions in policy evaluation, and is sometimes referred to as Goals-Based evaluation. In the American evaluation literature it is sometimes referred to as 'legislative monitoring', because it monitors whether the outcomes that were expected from some government policy initiative have been achieved. In the United Kingdom, the achievement of targets that have been set by Public Service Agreements and Service Delivery Agreements are evaluated using Goals-Based methods of evaluation.

An example in the UK context might be whether or not the goals and targets of the National Literacy Strategy (i.e. increasing the reading, writing and comprehension abilities of children and adults) have been achieved. Another example might be whether the goals of the Hospital Waiting Lists initiative (i.e. reducing the number of people on hospital waiting lists and/or the time they had to wait for treatment) have been achieved. Such outcomes may, or may not, be made explicit in policy statements and documents.

Goals Based evaluations make no assumptions about whether or not the chosen goals or targets are valid or appropriate measures of effectiveness. It may indeed be the case that waiting no more than four hours for hospital treatment is less valid to patients and their carers than waiting for two hours or less. Or it may be that waiting times for treatment are less valid than making sure that the most effective and evidence-based treatment methods are used by doctors and hospitals. Goals Based evaluations simply measure whether some goals or targets set by policy makers have been achieved.

Even when the goals of a policy, programme or project have been achieved, however, this does not necessarily mean that the policy in question has been responsible for this outcome. Other factors, including other policy initiatives, may have been responsible. In order to know whether the policy in question has been responsible for an anticipated outcome, some evaluation of the *counterfactual* is required (i.e. what would have happened anyway, or because of other interventions). Randomised control trial methods are generally considered to be the most appropriate way of determining the counterfactual of a policy, programme or project, though carefully controlled matched comparisons studies and some forms of statistical modelling also provide estimates of the counterfactual. These methods are reviewed in Why

do social experiments?, and further guidance on how statistical modelling can estimate the counterfactual can be found in The Green Book.

1.3.5 How do you evaluate unintended outcomes? – goals-free evaluation

Policy makers and evaluators are often interested in the *unintended* consequences or outcomes of a policy, programme or project. These unintended outcomes may be beneficial or harmful. *Goals-free* evaluation does this by focusing on the *actual* effects or outcomes of some policy, programme or project, without necessarily knowing what the intended goals might be. This type of policy evaluation is more commonly undertaken by evaluators who are independent of government and who are more interested in the *range* of consequences of a policy, programme or project than in the anticipated outcomes alone. Goals-free policy evaluation, however, should be of interest to government social researchers and policy analysts because of the importance of establishing the balance between the positive and negative consequences of policies. Such a balanced evaluation is important in order to establish the cost-benefit and cost-utility of a policy or programme intervention.

1.3.6 Experimental and quasi-experimental evaluation

Experimental and quasi-experimental research methods provide valid and reliable evidence about the *relative effectiveness* of a policy intervention compared with other policy interventions, or doing nothing at all (sometimes called the *counterfactual*). They provide appropriate evidence about questions such as whether a personal adviser service is more, or less, effective in terms of advancing low paid people in the labour market than, for example, providing skills training, or doing nothing at all.

The purest form of experimental method is the *randomised controlled trial* (RCT - sometimes called the *random allocation method* of evaluation). Randomised control trials deal with the problem of other possible factors influencing an outcome by exposing an experimental group of people, and a non-experimental (i.e. control) group of people to exactly the same factors *except* the policy, programme or project under investigation. The allocation of people to the experimental policy intervention, or to the control (i.e. no intervention) situation, is done purely on the basis of chance (i.e. randomisation). Randomisation does not guarantee that the experimental and control groups will be identical, but it reduces the influence of extraneous factors by ensuring that the only differences between the two groups will be those that arise by chance.

Randomisation may be by individuals or by units, clusters, or whole areas. Some welfare-to-work initiatives have allocated individuals (e.g. Job Seekers' Allowance claimants, people on Working Families Tax

Credits) to experimental or control groups. Other policy initiatives have allocated units such as schools, hospitals, housing estates or entire neighbourhoods, to experimental or control groups. The methods, problems and limitations of randomised controlled trials are discussed in Why do social experiments?.

Quasi-experimental methods refer to those research designs that compare the outcomes of experimental and control groups by methods other than randomisation. These include:

- controlled before and after designs involving pre-test and post-test comparisons using a single group of people (i.e. where individuals or units are their own controls).
- controlled before and after designs in which pre-test and post-test comparisons are made between two or more groups of people (i.e. experimental and external controls).
- interrupted time series studies (based on repeated observations over time of valid and reliable standardised measures of outcome).
- various types of matching designs using matched comparisons of individuals or units before and after an intervention.
- regression discontinuity designs.

Experimental and quasi-experimental designs are discussed in greater detail in Why do social experiments?.

1.3.7 Qualitative evaluation

Qualitative evaluations are designed to "permit the evaluator to study selected issues in depth and detail" Patton (1990). Such depth and detail is usually necessary to determine the appropriate questions to ask in an evaluation, and to identify the situational and contextual conditions under which a policy, programme or project works or fails to work.

Qualitative methods of evaluation are particularly important for formative evaluation which, as Patton (1990:156) again suggests, "is limited entirely to a focus on a specific context". Patton goes on to argue that:

> *"Formative evaluation services the purpose of improving a specific program, policy, group of staff (in a personnel evaluation), or product. Formative evaluations aim at 'forming' the thing being studied….There is no attempt in formative evaluation to generalise findings beyond the setting in which one is working. The purpose of the research is to improve effectiveness within that setting."*
>
> (Patton, 1990, 156)

Qualitative evaluation uses a range of methods including in-depth interviews, case studies, consultative methods, focus groups,

ethnography, observational and participant-observational studies, and conversation and discourse analysis. These methods of qualitative evaluation are discussed in greater detail in How do you know how (and why) something works? Qualitative methods of evaluation.

1.3.8 Economic appraisal and evaluation

Policies, programmes and projects involve the allocation of scarce and finite resources to competing demands and interests. The old adage that a pound cannot be spent twice means that choices between the competing demands upon a resource have to be made. Consequently, it is necessary to undertake economic appraisal at the outset (i.e. *ex ante*) of different policy options and the likely outcomes (both positive and negative) that will be achieved by them, and of the costs involved in achieving these outcomes. It is also necessary to undertake an economic evaluation after (i.e. *post hoc*) a chosen policy, programme and project has been running for some time in order to determine whether or not the anticipated outcomes (or other outcomes) have been achieved.

There are different types of economic appraisal and evaluation. The simplest type is cost appraisal and evaluation, which simply compares the costs of different initiatives without considering the outcomes to be achieved (or that have been achieved). The limitations of such appraisals and evaluations are fairly obvious – they tell us very little about the *relative effectiveness* or *benefits* of different interventions – and are of little value alone in policy evaluation.

Other types of economic appraisal and evaluation, which are more analytically powerful and useful to policy making, include *cost-effectiveness* and *cost-benefit* analyses. The former compares the differential costs involved in achieving a given objective, whereas the latter considers the differential benefits that can be gained by a given expenditure of resources. Cost benefit analysis involves a consideration of alternative uses of a given resource, or the *opportunity* cost of doing something compared with doing something else. Another type of economic appraisal is *cost utility* analysis, which evaluates the utility of different outcomes for different users or consumers of a policy or service. Cost utility analysis typically involves subjective appraisals and evaluations of outcomes using qualitative and quantitative data.

Economic appraisal and evaluation uses a variety of tools to estimate the costs and benefits of policy initiatives over time, such as the *discount rate* for adjusting the value of outcomes that will occur in the future. Detailed guidance on such tools, and on economic appraisal and evaluation more generally, are provided by The Green Book.

1.4 How does policy evaluation relate to project management?

Policy evaluation and analysis requires a structured and organised approach to defining an answerable question, summoning appropriate and relevant evidence, critically appraising and analysing that evidence, identifying the risks and opportunities of a policy, programme or project, and determining the likely effects (positive and negative) of the project at hand. Project and programme management has emerged in recent years as a structured and organised way of planning, implementing and concluding projects and programmes. The congruity of interest between policy evaluation and project management is clear.

1.5 Summary

Policy evaluation is a family of research methods that are used to systematically investigate the effectiveness of policies, programmes, projects and other types of social intervention, with the aim of achieving improvement in the social, economic and everyday conditions of people's lives. Different methods of policy evaluation are used to answer different questions. *The Magenta Book* provides a set of guidance notes on how to use the methods of policy evaluation and analysis effectively and, thereby, to generate and use sound evidence at the heart of policy making and implementation.

1.6 References

Chen, H.T. (1990) *Theory-Driven Evaluations*. Thousand Oaks, California, Sage Publications.

Cabinet Office (1999a) *Modernising Government*. White Paper. Cm 4310. London, HMSO. (http://www.policyhub.gov.uk/docs/modgov.pdf)

Cabinet Office (1999b) *Professional Policy Making for the 21st Century*. A report by the Strategic Policy Making Team. London, HMSO. (http://www.civilservant.org.uk/profpolicymaking.pdf)

Cabinet Office (2000) *Adding it up: improving analysis and modelling in Central Government*. A Performance and Innovation Unit Report. London, HMSO. (http://www.policyhub.gov.uk/docs/addingitup.pdf)

Bullock, H., Mountford, J., & Stanley, R. (2001) *Better Policy Making*. Centre for Management and Policy Studies, Cabinet Office, London. (http://www.civilservant.org.uk/betterpolicymaking.pdf)

Connell, J.P., Kubisch, A.C., Schorr, L.B. & Weiss, C.H. (1995) *New Approaches to Evaluating Community Initiatives: Concepts, Methods and Contexts*. Washington D.C., Aspen Institute.

Funnell, S. (1997) Program Logic: An Adaptable Tool, *Evaluation News and Comment*, 6(1): 5-17.

Greene, J.C., Benjamin. L, & Goodyear, L. (2001) The Merits of Mixing Methods in *Evaluation, Evaluation,* 7(1): 25-44.

HM Treasury (2003) *Green Book: appraisal and evaluation in Central Government* (http://www.hm-treasury.gov.uk/Economic_Data_and_Tools/Greenbook/data_greenbook_index.cfm)

Judge, J. & Bauld, L. (2001) Strong theory, flexible methods: evaluating complex community-based initiatives, *Critical Public Health*, 11(1): 19-38.

ODPM (2004) *Assessing the impact of spatial interventions: Regeneration, Renewal and Regional Development, 'The 3Rs guidance.'* Interdepartmental Group on the EGRUP Review, London (http://www.offpat.info/content-download/The%203Rs%20Guidance.pdf).

Owen, J.M. & Rodgers, P.J. (1999) *Program Evaluation: Forms and Approaches.* Thousand Oaks, California, Sage Publications.

Patton, M.Q. (1990) *Qualitative Evaluation and Research Methods*, 2nd edition. Newbury Park, California, Sage Publications.

Pawson, R. & Tilley, N. (1997) *Realistic Evaluation*. London, Sage.

Petrosino, A., Turpin-Petrosino, & Buehler, J. (2002) "*Scared Straight*" and *Other Juvenile Awareness Programs for Preventing Juvenile Delinquency.* Campbell Library. (http://www.campbellcollaboration.org/frontend2.asp?ID=4)

Rogers. P.J., Petrosino, A., Huebner, T.A., & Hacsi, T.A. (2000) *New Directions for Evaluation: Program Theory in Evaluation: Challenges and Opportunities.* San Francisco, Jossey Bass Publishers.

Rossi, P.H., Freeman, H.E. & Lipsey, M.W. (1999) *Evaluation: A systematic Approach*, 6th edition. Thousand Oaks, California, Sage Publications.

Scriven, M. (1991) *Evaluation Thesaurus*. 4th edition. Newbury Park, California, Sage Publications

Weiss, C.H. (1997) Theory-based evaluation: past, present and future, *New Directions for Evaluation*, 76: 41-55.

GOVERNMENT SOCIAL RESEARCH
Analysis for Policy

The Magenta Book: guidance notes for policy evaluation and analysis

Background paper 2: what do we already know? Harnessing existing research

Published July 2003
Updated October 2007

Government Social Research Unit

HM Treasury
1 Horse Guards Road
London SW1A 2HQ

2 What do we already know?

2.1 The problem

An essential first step in planning a policy evaluation is to determine what is already known about the topic in question from the full range of existing evidence. This is important for at least four reasons:

1. It may be that there is already sufficient evidence on the likely effectiveness of a policy, programme or project so that further primary evaluation is unnecessary. Such a situation is very unlikely for the reasons outlined below.

2. The existing evidence may be ambiguous, inconclusive, or of uncertain quality indicating that further evaluation is necessary and that specific aspects of the policy in question need addressing.

3. It may be that there is no valid, reliable and relevant evidence available at all on the policy in question. This will help determine the nature and scope of the evaluation that needs to be undertaken.

4. Any single evaluative study may illuminate only one part of a policy issue, or its findings may be sample specific, time specific, or context specific. This makes it difficult to establish the generalisability and transferability of findings from existing research evidence which, in turn, will influence what requires evaluating.

Establishing what is already known about a policy, programme or project, however, presents a major challenge for knowledge management. The sheer amount of potential research evidence in most substantive areas of social science and public policy, coupled with the rapid growth of access to knowledge and information as a result of information technology, make it almost impossible to keep abreast of the research literature in any one area. Given the limitations of humans' information processing abilities, the complexity of modern professional life almost certainly exceeds the capacity of the *unaided* human mind (Eddy 1999).

The problems of information overload are compounded by the fact that not all research and information is of equal value. Variations in the quality of primary studies, reporting practices, standards of journal indexing and editing, and publication criteria mean that the existing research literature is often of variable quality. Consequently, seemingly similar studies may be of different focus, value and relevance to users of research evidence. Some way of differentiating between high and lower quality studies, as well as relevant and irrelevant evidence, is required.

2.2 A solution

Systematic reviews of existing research literature are increasingly being used as a valid and reliable means of harnessing the existing research evidence. They can also allow a cumulative view of existing research evidence to be established. As Cooper and Hedges (1994:4) point out, systematic reviews "attempt to discover the consistencies and account for the variability in similar-appearing studies". Also, "seeking generalisations also involves seeking the limits and modifiers of generalisations" (*ibid*) and, thereby, identifying the contextual-specificity of available research evidence.

2.3 How is this different from what is normally done?

Systematic reviews differ from other types of research synthesis (e.g. narrative reviews and vote counting reviews) by:

- being more systematic and rigorous in the ways they search and find existing evidence;
- having explicit and transparent criteria for appraising the quality of existing research evidence, especially identifying and controlling for different types of bias in existing studies;
- having explicit ways of establishing the comparability (or incomparability) of different studies and, thereby, of combining and establishing a cumulative view of what the existing evidence is telling us.

Two common methods of synthesising existing evidence are *narrative reviews* and *vote counting reviews.*

2.3.1 Narrative reviews

The simplest form of research synthesis is the traditional qualitative literature review, often referred to as the *narrative review.* Narrative reviews typically attempt to identify:

- readily available literature on a subject or topic;
- which methodologies have been used in that literature;
- what samples or populations have been studied (and not studied);
- what findings have been established;
- what caveats, qualifications and limitations exist in the available literature.

Narrative reviews may (or may not) provide an overview or summary of research on a topic. More typically they identify the range and diversity

of the available literature, much of which will be inconsistent or inconclusive.

A major limitation of narrative reviews is that they are almost always *selective*. They do not always involve a *systematic, rigorous* and *exhaustive* search of *all* the relevant literature using electronic and print sources as well as hand searching and ways of identifying the 'grey' literature (i.e unpublished studies or work in progress). This means that traditional narrative literature reviews often involve *selection bias* and/or *publication bias*. The latter is a consequence of some journals disproportionately reporting studies with positive outcomes, whilst some other sources disproportionately report studies with negative outcomes.

Narrative literature reviews are also often *opportunistic* in that they review only literature and evidence that is readily available to the researcher (the file drawer phenomenon). Some narrative reviews may discard studies that use methodologies in which the researcher has little or no interest. Alternatively, they may include studies that use different methodologies and which do not lend themselves to meaningful comparison or aggregation. Narrative reviews often provide few details of the procedures by which the reviewed literature has been identified and appraised. It is also often unclear how the conclusions of narrative reviews follow from the evidence presented. This lack of transparency makes it difficult to determine the selection bias and publication bias of narrative reviews, and runs the risk of over-estimating (or in some cases under-estimating) the effectiveness of interventions in ways that are hard to identify.

With systematic reviews the problems of selection bias and publication bias are dealt with by identifying and critically appraising *all* of the available research literature, published and unpublished. This involves detailed hand searching of journals, textbooks, and conference proceedings, as well as exhaustive electronic searching of the existing research literature.

Systematic reviews also differ from narrative reviews in that they make explicit the search procedures for identifying the available literature, and the procedures by which this literature is critically appraised and interpreted. This affords a degree of transparency by which other researchers, readers and users of systematic reviews can determine what evidence has been reviewed, how it has been critically appraised, and how it has been interpreted and presented. This, in turn, allows for other interpretations of the evidence to be generated, and for additional studies of comparable quality to be added to the review, if and when they become available. In these ways, an interactive and cumulative body of sound evidence can be developed.

2.3.2 Vote counting reviews

A type of research synthesis that attempts to be cumulative is the *vote counting review*. This attempts to accumulate the results of a collection

of relevant studies by counting "how many results are statistically significant in one direction, how many are neutral (i.e. "no effect"), and how many are statistically significant in the other direction" (Cook *et al,* 1992:4). The category that has the most counts, or votes, is taken to represent the modal or typical finding, thereby indicating the most effective means of intervention.

An obvious problem with voting counting reviews is that they do not take into account the fact that some studies are methodologically superior than others and, consequently, deserve special weighting. Systematic reviews differentiate between studies of greater and lesser sample size, power and precision and weight them accordingly. (To see how such weighting of different studies is done see Deeks, Altman and Bradburn, (2001)).

Another problem with vote counting reviews is that they fail to indicate "the possibility that a treatment might have different consequences under different conditions" (Cook *et al*, 1992:4). Crude counting of studies in terms of the direction of outcomes does not take into account that "person and setting factors are especially likely to moderate causal relationships and help explain why a treatment has the effects it does" (Cook *et al*, 1992:22). Systematic reviews attempt to incorporate such contextual factors by closely analysing the findings and limitations of different studies and identifying their implications for policy and practice. Where there is evidence on a topic from qualitative research this can also be used to identify important contextual and mediating factors.

2.4 What is meta-analysis?

Meta-analysis is a type of systematic review that aggregates the findings of comparable studies and "combines the individual study treatment effects into a "pooled" treatment effect for all studies combined" (Morton, 1999). The term 'meta-analysis' has been commonly attributed to Gene Glass (1976) who used the term to refer to "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings". The statistical basis of meta-analysis, however, can be traced back to seventeenth century astronomy, which suggested "that combinations of data might be better than attempts to choose amongst them" (Egger, Davey-Smith and O'Rourke, 2001:8).

In the two decades or more since Glass's original meta-analytic work on psychotherapy (Smith, Glass and Miller, 1980) and class size (Glass and Smith, 1979; Smith and Glass, 1980; Glass, Cahen, Smith and Filby, 1982), meta-analysis has developed considerably in terms of the range and sophistication of data-pooling and statistical analysis of independent studies (see Kulik and Kulik, 1989, Cook *et al*, 1992, Cooper and Hedges, 1994 and Egger, Davey Smith, and Altman, 2001 for more detailed accounts of these developments). They have also been

undertaken in substantive areas other than education and health care, including criminology, social work and social welfare.

Meta-analysis is perhaps best known for combining the results of randomised controlled trials (see Why Do Social Experiments?), though as Egger, Davey-Smith and Schneider (2001:211) point out they are also commonly undertaken on non-randomised data from primary studies that use case-control, cross-sectional, and cohort designs. Non-randomised studies, however, are much more susceptible to the influence of confounding factors and bias and may "provide spuriously precise, but biased, estimates of association" (*ibid*).

Meta-analysis of randomised controlled trials, on the other hand, assumes that each individual trial provides an unbiased estimate of the effects of an experimental intervention, and that any variability of results between studies can be attributed to random variation. Consequently, by combining the results of randomised controlled trials an overall effect of the intervention can be estimated that is unbiased and has measurable precision.

2.5 What are the limitations of meta-analysis?

2.5.1 The "apples and pears" problem

Meta-analysis has its own limitations. Like other types of research synthesis it requires *focussed* questions to be asked about:

- the intervention(s) under investigation
- the population (or sub-groups) studied
- the outcomes that are being assessed.

Given that each of these factors may vary across individual studies this presents a challenge for the meta-analyst to ensure that there is real consistency between primary studies on all three dimensions. If this is not done there is the "apples and pears" problem of falsely aggregating studies that are not really comparable.

There are ways of testing whether or not different primary studies are sufficiently similar (or homogeneous) for their findings to be aggregated into a pooled estimate of overall effect size. Funnel plot analysis is one such method of testing for homogeneity or heterogeneity of different primary studies (see Deeks, Altman and Bradburn, 2001). Moreover, there are different ways of analysing studies where there is greater homogeneity (i.e. using fixed effects models) and where there is greater heterogeneity (i.e. using random effects models). For further discussion of random-effects and fixed-effects models of meta-analysis see Hedges (1994), Raudenbusch (1994) and Deeks, Altman and Bradburn, (2001).

2.5.2 The adequacy of searching

Meta-analysis may also be limited by the degree of systematic and comprehensive searching that is undertaken for relevant and appropriate primary studies. Extensive, if not exhaustive, searches are required of databases, textbooks, journals, conference proceedings, dissertation abstracts, and research-in-progress, using electronic and hand searching methods. The need to search unpublished sources (including research-in-progress) is crucial given the problems of positive (and in some cases negative) publication bias in journals and other print sources.

2.5.3 The quality of primary studies

Meta-analysis requires high quality standards of methodology and reporting in the primary studies being synthesised. These include:

- the validity and reliability of tests and outcome measures used in the primary studies;
- follow-up data that are consistent with baseline data;
- the reporting of (and accounting for) participants lost-to-follow-up at different data collection points (i.e. attrition bias);
- accounting for missing data on moderator and mediating variables;
- systematic differences in the treatment of comparison groups other than the intervention under investigation (i.e. performance bias);
- the appropriateness of descriptive and inferential statistics (means and standard deviations, chi-squares, odds ratio and confidence intervals) used in the primary studies;
- the types of statistical manipulation (e.g. logarithmic transformations of data) used in the primary studies;
- ensuring the independence of primary studies so that the results of individual studies are not included more than once in a meta-analysis, thereby double counting studies in estimating the effect size.

As Cook *et al* (1992) point out, there are several ways in which problems of inadequate statistical reporting can be handled by meta-analysts. These include:

- using external sources to establish the validity and reliability of instruments used in primary studies;
- contacting the primary investigator(s) to obtain additional data or clarification of procedures used;

- reporting deficiencies of primary data in the meta-analysis, thereby distinguishing between good and poor data.

All of these problems need to be confronted and resolved by meta-analysts in order to provide unbiased estimates of the overall likely effects of an intervention and greater precision than that given by narrative or vote counting reviews.

2.6 What about the synthesis of non-experimental studies?

Methods for synthesising data from primary studies that use experimental (randomised control trials) and quasi-experimental methods (such as case-control, cross-sectional, and cohort designs) are well developed. The synthesis of primary studies that use non-experimental methods such as in-depth interviews, observational methods, participant observation and ethnography is currently less developed. Strategy Unit in the Cabinet Office commissioned the National Centre for Social Research to undertake a methodological review of quality standards in qualitative evaluation methods, and to develop a framework for the critical appraisal of qualitative evaluation studies. Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence (Spencer *et al*, 2003 on behalf of the Cabinet Office). Procedures for undertaking systematic reviews of different types of evidence have been developed (EPPI-Centre, 2001), as have methods for synthesising qualitative research (see for example Thomas and Harden (2007) Methods for the thematic synthesis of qualitative research in systematic reviews. For further information on synthesing findings see the EPPI-Centre.

Two earlier attempts to develop the synthesis of non-experimental studies include *meta-ethnography* and *best evidence synthesis.*

2.6.1 Meta-ethnography

Meta-ethnography attempts to summarise and synthesise the findings of qualitative studies, especially ethnographies and interpretive studies. Meta-ethnography claims to "be interpretive rather than aggregative" (Noblit and Hare, 1988:11) and covers:

> *"research that is termed ethnographic, interactive, qualitative, naturalistic, hermaneutic, or phenomenological. All these types of research are interpretive in that they seek an explanation for social or cultural events based upon the perspectives and experiences of the people being studied. In this way, all interpretive research is "grounded" in the everyday lives of people."*
>
> (Noblit and Hare, 1988:12)

Like meta-analysis, meta-ethnography "seeks to go beyond single accounts" (Noblit and Hare, 1988:13), but instead of doing so by

aggregating samples and identifying consistencies and variability between different studies, it does this by "constructing interpretations, not analyses" and by revealing "the analogies between the accounts" (*ibid*).

Meta-ethnography "reduces the accounts while preserving the sense of the account through the selection of key metaphors and organisers" (*ibid*). These refer to "what others may call themes, perspectives, organisers, and/or concepts revealed by qualitative studies" (*op cit:* 14). To this extent, meta-ethnography would appear to have more in common with narrative reviews than with vote counting reviews and meta-analyses.

2.6.2 What are the problems and limitations of meta-ethnography?

Meta-ethnography has some of the same problems as meta-analysis and other types of research synthesis. These include:

- establishing criteria for which studies to include and exclude in a meta-ethnographic review;
- handling the diversity of the questions being asked and the theoretical perspectives from which these questions are generated;
- dealing with the heterogeneity of primary studies that use qualitative research and evaluation methods;
- balancing summary statements of qualitative studies with their contextual specificity;
- providing overviews of qualitative studies without ignoring the "meaning in context" and the "ethnographic uniqueness" that is central to ethnographic and qualitative inquiry.

Meta-ethnography is seen by those who require quantitative synthesis of existing evidence as being limited by its inability:

- to provide statistical accumulation of findings;
- to allow prediction or to specify any degree of confidence about qualitative findings;
- to allow for the statistical control of bias;
- to test for, and control, the heterogeneity/homogeneity of different studies.

These latter concerns about the synthesis of qualitative studies, however, seem to miss the point of what ethnographies and other qualitative studies are trying to achieve (Davies, 2000). That is, to provide rich descriptions of naturally occurring activity, rather than experimental constructions, and to interpret the individual and shared

meanings of the topic under investigation for different individuals and groups. Qualitative inquiry is particularly concerned with the contextual specificity of these meanings rather than with their de-contextualised generalisability. To subject such findings to statistical representation, manipulation and testing is usually inappropriate, other than to identify patterns of consistency and inconsistency amongst different qualitative studies.

2.6.3 Best evidence synthesis

Slavin (1984, 1986) has proposed that the type of methods used to generate research evidence is less important than the quality of the primary studies undertaken, whatever methodological approaches are used. Slavin suggests that what is required is 'best evidence synthesis' in which "reviewers apply consistent, well justified, and clearly stated a *priori* inclusion criteria" of studies to be reviewed. Primary studies should be "germane to the issue at hand, should be based on a study design that minimises bias, and should have external validity". The latter requires outcome variables that have some 'real life' significance rather than "extremely brief laboratory studies or other highly artificial experiments" (*ibid*).

More recently Slavin and Fashola (1998) have presented a best evidence synthesis of "proven and promising programs for America's schools", which uses this rather pragmatic notion of research synthesis. Some studies are included in this review even though Slavin and Fashola had reservations about some aspects of the primary studies in question. They note, for instance, that the comparison groups used in Mehan *et al's* (1996) AVID project may be susceptible to bias, yet they conclude that "the college enrolment rates for AVID are impressive, and the program has a good track record in serving students throughout the United States. The Mehan *et al* study provides good qualitative evidence from case studies, interviews with students and teachers, and ethnographic research, of *why* and *how* the AVID programme succeeds, and has limitations. For these reasons, say Slavin and Fashola, this study is "worthy of consideration by other schools serving many students placed at risk" (Slavin and Fashola, 1998:87).

2.7 What relevance does all this have for government research and evaluation?

Evidence-based principles are at the heart of the Government's reform agenda for better policy making and policy implementation. "What matters is what works" is a repeated theme of government policy documents and Ministerial statements. Consequently, it is essential that Government departments and agencies have ways of accessing, harnessing and using the best available research evidence for effective policy making.

One of the frequent criticisms of systematic reviews for Government purposes is that they take a long time to complete (between six months and one year), and that potential Government users of reviews require evidence more rapidly. Establishing an evidence-base in any subject does take time, and building up a body of sound evidence is a lengthy process. Users of research and evaluation evidence often need quicker access to what the existing evidence is telling them, and what gaps remain in the research evidence on some topic or question.

2.7.1 Rapid evidence assessments

To this end, Rapid Evidence Assessments are being developed for use in public policy research and evaluation. Rapid Evidence Assessments are appraisals of existing evidence that sit somewhere between *the equivalent* of Health Technology Assessments (HTAs) and fully developed systematic reviews in the field of health care.

HTAs are descriptive rather than analytical abstracts of healthcare interventions that have not been critically appraised and fully evaluated according to systematic review procedures. Nonetheless, they include "evidence of clinical outcomes relative to no treatment and/or the best existing treatment for the condition in question, including undesirable side-effects and, (for chronic conditions) effects of stopping treatment" (NHS Executive, 1999). In addition, HTAs include estimates of:

- the impact on quality and length of life;
- estimates of the average health improvement per treatment initiated;
- net NHS costs associated with this health gain;
- other (non-NHS) costs and savings caused by the intervention;
- any significant differences between patients and sub-groups of the population;
- the expected total impact on NHS resources (including manpower resources).

HTAs typically take between 8 and 12 weeks to assemble.

Whilst other areas of policy and practice differ in some respects from health care there are parallels that are worth developing in terms of generating structured appraisals of what works, how, for whom, with what potential negative effects, and at what costs and benefits. Rapid Evidence Assessments will collate descriptive outlines of the available evidence on a topic, critically appraise them (including an economic appraisal), sift out studies of poor quality, and will provide an overview of what that evidence is telling us, and what is missing from it. They will be based on fairly comprehensive electronic searches of appropriate databases, and some searching of print materials, but not the exhaustive

database searching, hand searching of journals and textbooks, or searches of the grey literature that go into systematic reviews.

It is anticipated that Rapid Evidence Assessment will be completed and available in less than 8-12 weeks, though this will depend on the topic under investigation, the available evidence, and the available resources to review, appraise and summarise the evidence. Rapid Evidence Assessments will carry a caveat that their conclusions may be subject to revision once the more systematic and comprehensive review of the evidence has been completed. This is consistent with the important principle that systematic reviews are only as good as their most recent updating and revision allows.

2.8 Where can I find help with systematic reviews and harnessing existing evidence?

There are a number of academic and government agencies that provide advice, guidance and specialist expertise on how to develop, and use, systematic reviews of research evidence. Some of these agencies undertake the preparation and dissemination of systematic reviews and other types of research synthesis.

The Campbell Collaboration (http://www.campbellcollaboration.org/) is an international network of social scientists and policy analysts that prepares, maintains and disseminates systematic reviews of the effectiveness of interventions in education, crime and justice, and social welfare. It also provides methodological guidance and some training on how to undertake systematic reviews, and quality assurance procedures for generating valid and reliable reviews. Research and evaluation groups from around the world contribute to the Campbell Collaboration.

The Cochrane Collaboration (http://www.cochrane.org/index.htm) is the forerunner of the Campbell Collaboration and prepares, maintains and disseminates systematic reviews of the effects of interventions in health care. The Cochrane Collaboration has an impressive electronic library of systematic reviews in over 50 areas of medicine and health care. It also has nine methods groups and provides informative guidance on the methodology of systematic reviews. The Cochrane Reviewers' Handbook (available via the above website address) is a valuable source of guidance on how to undertake and appraise systematic reviews.

The Department for Education and Skills (DfES) has commissioned a Centre for Evidence-Informed Policy and Practice in Education (the EPPI-Centre), which is located at the Institute of Education at the University of London (http://eppi.ioe.ac.uk/). The EPPI Centre undertakes and commissions systematic reviews in education, and is developing methods for undertaking systematic reviews of social science and public policy research.

The Economic and Social Research Council (ESRC) has established an Evidence Network, which consists of a Centre for Evidence-Based Policy and Practice at Queen Mary College, London and seven evidence 'nodes'. The Centre for EBPP is also developing the methodology of systematic reviews in the social sciences and public policy field, and is establishing a database of high quality reviews. Further details are available at http://www.evidencenetwork.org/.

There are some very useful textbooks and handbooks on systematic reviews and research synthesis. Sources which deserve special mention are:

Cochrane Handbook for Systematic Reviews of Interventions
http://www.cochrane.org/resources/handbook

Cook, T.D., Cooper, H., Cordray, D.S., Hartmann, H., Light, R.J., Louis, T.A. and Mosteller, F., (1992) *Meta-Analysis for Explanation.* New York, Russell Sage Foundation.

Cooper, H. and Hedges, L.V. (eds), (1994) *The Handbook of Research Synthesis*. New York, Russell Sage Foundation.

Egger, M., Davey Smith, G. and Altman, D.G. (eds), (2001) Systematic Reviews. In *Health Care: Meta-Analysis in Context*. London, BMJ Publishing Group.

Slavin, R.E. and Fashola, O. S., (1998) *Show Me the Evidence!* : *Proven and Promising Programs for American Schools.* Thousand Oaks, California, Corwin Press.

An Analysts' Checklist for Undertaking a Systematic Review is presented at Annexe A below.

A Policy Makers' Checklist for Using a Systematic Review is presented at Annexe B below.

2.9 Conclusion

There is a growing recognition of the potential of systematic reviews and other types of research synthesis for policy evaluation and analysis. Systematic reviews provide a powerful way of harnessing existing evidence that is valid, reliable and transparent. They differ from traditional narrative reviews and other types of literature review in that they use exhaustive methods for searching evidence, critically appraise that evidence according to explicit criteria, and identify the implications for policy and practice only from research that has passed high quality control procedures. Systematic reviews are not a panacea, nor are they a substitute for sound judgement and expert decision making. Rather, they provide one means of establishing a sound empirical research base upon which the judgements and expertise of decision makers can be made.

2.10 References

Cook, T.D., Cooper, H., Cordray, D.S., Hartmann, H., Light, R.J., Louis, T.A. & Mosteller, F. (1992) *Meta-Analysis for Explanation.* New York, Russell Sage Foundation.

Cooper, H. & Hedges, L.V. (eds) (1994) *The Handbook of Research Synthesis.* New York, Russell Sage Foundation.

Davies, P.T. (2000) Contributions from qualitative research. In Davies, H.T.O., Nutley, S.M. & Smith, P.C. (eds) *What Works? Evidence-based Policy and Practice in Public Services*. Bristol, The Policy Press.

Deeks, J.J., Altman, D.G. & Bradburn, M.J. (2001) Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In Egger, M., Davey Smith, G. & Altman, D.G. (eds) *Systematic Reviews in Health Care: Meta-Analysis in Context*. London, BMJ Publishing Group.

Eddy, D. (1999) *Doctors, Economics and Clinical Practice Guidelines: Can they Be Brought Together?*. London, Office of Health Economics.

Egger, M., Davey Smith, G., & O'Rourke, K. (2001) Rationale, potentials, and promise of systematic reviews. In Egger, M., Davey Smith, G. & Altman, D.G. (eds) *Systematic Reviews in Health Care: Meta-Analysis in Context*. London, BMJ Publishing Group.

Egger, M., Davey Smith, G. & Altman, D.G. (eds) (2001) *Systematic Reviews in Health Care: Meta-Analysis in Context*. London, BMJ Publishing Group.

Egger, M., Davey Smith, G., & Schneider, M. (2001) Systematic reviews of observational studies. In Egger, M., Davey Smith, G. & Altman, D.G. (eds) *Systematic Reviews in Health Care: Meta-Analysis in Context*.. London, BMJ Publishing Group.

EPPI-Centre, 2001 *Review Group Manual, Version 1.1*. London, Institute of Education.

Glass, G.V. (1976) Primary, secondary and meta-analysis of research, *Educational Researcher*. 5: 3-8.

Glass, G.V. & Smith, M.L. (1979) Meta-analysis of research on class size and achievement, *Educational Evaluation and Policy Analysis*, 1: 2-16.

Glass, G.V., Cahen, L.S., Smith, M.L., & Filby, N.N. (1982) *School Class Size: Research and Policy*. Beverley Hills, Sage Publications.

Thomas, J. Harden, A. (2007) *Methods for the thematic synthesis of qualitative research in systematic reviews.* NCRM Working Paper Series Number (10/07).

Hedges, L.V. (1994) Fixed Effects Models. In Cooper, H. & Hedges, L.V. (eds) *The Handbook of Research Synthesis*. New York, Russell Sage Foundation.

Kulik, J.A. & Kulik, C-L. C. (1989) Meta-analysis in education, *International Journal of Educational Research*, 13: 221-340.

Mehan, H., Villanueva, T., Hubbard, L. & Lintz, A. (1996) *Constructing School Success: The Consequences of Untracking Low-Achieving Students*. Cambridge, Cambridge University Press.

Morton, S. (1999) *Systematic Reviews and Meta-Analysis, Workshop materials on Evidence-Based Health Care*. University of California, San Diego, La Jolla, California, Extended Studies and Public Programs.

Noblit, G.W. & Hare, R.D. (1988) *Meta-Ethnography: Synthesizing Qualitative Studies*. Newbury Park, Sage Publications.

NHS Executive (1999) *Faster Access to Modern Treatment: How NICE Appraisal Will Works*. Leeds, National Health Service Executive.

Raudenbusch, S.W. (1994) Random Effects Models. In Cooper, H. & Hedges, L.V. (eds) *The Handbook of Research Synthesis*. New York, Russell Sage Foundation.

Slavin, R. E. (1984) Meta-analysis in education: How has it been used?, *Educational Researcher*, 13: 6-15.

Slavin, R.E. (1986) Best evidence synthesis: An alternative to meta-analysis and traditional reviews, *Educational Researcher*, 15: 5-11.

Slavin, R.E. & Fashola, O. S. (1998) *Show Me the Evidence!: Proven and Promising Programs for American Schools*. Thousand Oaks, California, Corwin Press.

Smith, M.L. & Glass, G.V. (1980) Meta-analysis of research on class size and its relationship to attitudes and instruction, *American Educational Research Journal*, 17: 419-433.

Smith, M.L., Glass, G.V. & Miller, T.I. (1980) *The Benefits of Psychotherapy*. Baltimore, Johns Hopkins University Press.

Spencer, L., Ritchie, J., Dillon, L. National Centre for Social Research (2003) *Quality in Qualitative Evaluation* produced on behalf of the Cabinet Office.

Petrosino, A.J., Boruch, R. Rounding, C., McDonald, S., & Chalmers, I. (2002) *A Social, Psychological, Educational & Criminological Trials Register (SPECTR).* Philadelphia, Campbell Collaboration (http://geb9101.gse.upenn.edu/).

Annexe A

Analysts' checklist for undertaking a systematic review

This checklist should only be used in conjunction with one or more of the existing guides to systematic reviews mentioned in paragraph 8.6 of these guidance notes.

Analysts proposing to undertake a systematic review for the first time are also advised to take a structured course on the topic, such as the Government Social Research Unit course on 'Methods for Synthesis'.

Formulating an answerable question

Does the central question of the review clearly address the following points?

- ❑ The policy *intervention* for which evidence is sought
- ❑ The *population* or *sub-groups* that the policy is expected to effect
- ❑ The *outcomes* that the policy intervention is expected to achieve

Searching for relevant studies

Have the following steps of a search strategy been planned?

- ❑ The searching of appropriate *electronic/internet* sources
- ❑ The searching of appropriate *print* sources (e.g. journals, textbooks, research reports)
- ❑ The *hand searching* of appropriate print sources
- ❑ Searching of the '*grey*' (i.e. unpublished) literature

Critically appraising studies found

How will the existing literature be sifted for quality and validity?

- ❑ The *appropriateness* of the questions, populations and outcomes addressed
- ❑ Evidence of *selection bias* in the primary studies
- ❑ Evidence of *performance bias* in the primary studies
- ❑ Evidence of *attrition bias* in the primary studies
- ❑ Evidence of *detection bias* in the primary studies

- What criteria will be used for *including and excluding* primary studies

Extracting data from included studies

Has a strategy been planned for extracting data from the included studies?

- A data collection form recording how, and why, data were extracted from included studies
- Information about the characteristic of included studies
- Verification of study eligibility for the review
- Details of study characteristics
- Details of study methods
- Details of study participants (populations and sub-groups)
- Details of study interventions
- Details of study outcomes and findings
- Reliability check for data collection/extraction

Analysing and presenting the findings

- What comparisons should be made (e.g. by interventions studied, participants included, outcomes measured)?
- What study results are needed for each comparison?
- What assessments of validity are to be used in the analysis?
- Is any other data or information needed from authors of studies included in the review?
- Do the data from different studies need to be transformed for the review's analysis?
- How is heterogeneity/homogeneity of studies to be determined?
- Is a meta-analysis of findings possible?
- What are the main findings of the review?
- What are the likely effect sizes of the proposed policy intervention, net of the counterfactual?
- What are the main caveats and/or qualifications of the findings of this review?

Interpreting the findings

- What is the strength of the evidence from the review?
- How applicable are the results of the review to 'real life' policy and practice?
- What does the review say about the costs and benefits of the proposed intervention?
- What trade-offs are suggested by the review between expected benefits, harm and costs (including opportunity costs)?
- What mediating factors emerge from the review that might affect the implications for policy and practice e in different contexts?

Summarising the implications for policy and practice

- What are the 'take home' messages for policy making and/or practice?
- What are the 'take home' messages for future research in this area?

Annexe B

B.1 Policy makers' checklist for using a systematic review

The following are suggested questions that policy makers should ask when reading a systematic review.

B.I.I The question you want answered

Has the reviewed covered the following features?

- ❑ The policy intervention for which evidence is sought
- ❑ The population or sub-groups that the policy is expected to effect
- ❑ The outcomes that the policy intervention is expected to achieve

B.I.II The adequacy of the review

Are there sufficient details in the review about the search strategy used to find relevant research evidence? i.e.

- ❑ Were appropriate electronic/internet sources searched?
- ❑ Were appropriate print sources (e.g. journals, textbooks, research reports) searched?
- ❑ Was any hand searching of appropriate print sources used?
- ❑ Was an attempt made to identify the 'grey' (i.e. unpublished) literature

B.I.III Critical appraisal of evidence

- ❑ Were the following tasks undertaken in the review?
- ❑ Was the research evidence that was identified sifted and graded for quality?
- ❑ Were the inclusion and exclusion criteria of primary studies made explicit?
- ❑ Were the appropriate outcome measures included in the review?

B.I.IV Quality of evidence presented

- ❑ Is the research evidence presented in the review easy to understand?

- Is a full evidence table presented?
- Has the strength of the existing evidence been assessed?
- Are there any estimates of the likely effect size of the policy intervention?
- Are there any details about the contexts in which the policy is likely to be effective?
- Is there any information on the likely costs and benefits of the proposed policy?

B.I.V Applicability of the evidence

- Has the evidence been presented in a way that is helpful to decision making?
- What are the implications of the review for policy and practice in this area?
- Are there any mediating factors from the review that need to be taken into account?
- What are the implications of this review for future research and evaluation in this area?

B.I.VI Peer review

- Has this systematic review been peer reviewed by independent analysts with expertise and experience in research synthesis?
- If not, do you want to get this systematic review peer reviewed?

The Magenta Book: guidance notes for policy evaluation and analysis

3 Theory based and realist approaches to evaluation

Paper forthcoming

Government Social Research Unit

HM Treasury
1 Horse Guards Road
London SW1A 2HQ

GOVERNMENT SOCIAL RESEARCH
Analysis for Policy

The Magenta Book: guidance notes for policy evaluation and analysis

Background paper 4: what do the statistics tell me? Statistical concepts, inference & analysis

Published April 2004
Updated October 2007

Government Social Research Unit

HM Treasury
1 Horse Guards Road
London SW1A 2HQ

4 What do the statistics tell me?

4.1 Introduction

This chapter is intended as an introduction to quantitative research. It first describes a range of methods that can be used to summarise and interpret statistical data. This is then followed by a description of more complex analytical methods, such as hypothesis testing and regression analysis. The reader though should recognise that there are a range of other advanced statistical methods that are not covered in this chapter (i.e. factor analysis, multi-level modelling, ANOVA etc).

4.2 Descriptive statistics

Data in their raw form usually consist of many rows of information and it is therefore not possible to obtain any useful conclusions from simply inspecting the raw data. It is one of the tasks of the data analyst to make meaningful sense of such data, by employing techniques that summarise the data and show relationships within the data.

In order to describe data succinctly, a range of descriptive statistics (i.e. summary measures) are used to investigate and present them. As an example, if one were reporting the heights of adult men from the 2001 Health Survey for England (HSE) (Bajekal et al, 2003), then one approach could be to simply list all the heights of all the men in the survey. However, given that the height was recorded for 6,542 men, such a list would not give the reader anything more than a vague sense of the range of heights measured. The data are therefore summarised, most often using a measure of their *centre* (e.g. the mean) and a measure of their *spread*, usually the variance.

4.3 Inferential statistics

Inferential data analysis involves exploring and determining relationships between variables (or sub-groups). Rarely are we just interested in the relationship within the sample, but more often we want to know if such relationships can be generalised to the whole population. Inferential data analyses involve trying to answer questions, such as, "do men and women have the same annual income on average?" or, "does smoking cause lung cancer?", etc.

4.4 Levels of measurement

Conventionally, variables can be measured at four different levels of measurement although in practice social scientists use only three of them (see sections 4.4.1 to 4.4.3).

4.4.1 Nominal variables

These are variables where numerical codes are used to identify groups or categories and the variable measures only differences between cases. For example, a variable measuring political preference is measured at the nominal level. If one person is recorded as voting Labour and another Conservative, this means only that they vote differently. This remains true even if, for the purpose of the analysis, Labour voters are coded 1 and Conservative voters as 2. Although Tory voters are coded 2, one cannot conclude that they are twice as political as Labour voters coded 1. The numbers are used only as indications that the respondents are in different categories. With nominal data, we can interchange codes without loss of information (1=Conservative, 2=Labour).

4.4.2 Ordinal variables

The numerical codes in ordinal variables have some order or rank in terms of bigger and smaller. For example, we may code level of education such that 1=no qualification; 2=O-levels, 3=A-levels, 4=degree. In this case, the values 1-4 have some ranking in the sense that someone who has a degree is clearly better educated than someone who only has O-levels. However, apart from the ranking, we cannot say that those with a degree (code 4) are four times as well educated as those with no qualifications (code 1). Variables measured at the nominal and ordinal levels are sometimes referred to as *categorical*.

4.4.3 Interval and ratio variables

These types of variables are measured on a *continuous* scale where the values are real numbers. Interval data have an *arbitrary zero point* while ratio-scale data have a *true zero*. The classic example of interval data is the Fahrenheit temperature. If substance X has temperature 100°F and Z=50°F, we know that X is hotter than Z by an interval of 50°F. However, we cannot say that X is *twice as hot* as Z. This is because 0°F does not mean 'no temperature'. For ratio scale, on the other hand, if X=6kg, and Z=2kg, then we can say that X is three times as heavy as Z. This is because 0kg means lack of (relative) weight. Most continuous variables in the social sciences are ratio-scale variables.

4.5 Measures of central tendency

There are several reasons for using a measure of the centre of a set of values to summarise its location rather than, for example, the maximum value. The value of the *centre* is a more stable measure under different sample sizes - it will vary at random within a fairly small range. (Whereas the maximum value for example would be likely to be greater for larger samples as there is a higher chance of observing a more extreme value.) It is also easier to interpret the measure of the *centre* as the value for an 'average' member of the sample. For these and other reasons, the *centre* is considered most appropriate as an 'anchoring point' for the values.

A range of measures is available to identify the *centre* (or average) of a set of values (see sections 4.5.1. to 4.5.3).

4.5.1 The arithmetic mean

The most common measure is the arithmetic mean, partly because it has useful statistical properties. The mean is calculated as the sum of all the values divided by the number of values. So, as a simple example, the mean of the values (2,2,4,7,10) is:

$$mean = \frac{2+2+4+7+10}{5} = \frac{25}{5} = 5.$$

4.5.2 The median

The median is the value which is at the *middle* of the distribution and is defined so that half the observations are smaller, and half are larger than it. As described above, the median is relatively unaffected by extreme values and thus suited as a measure of the 'average' for heavily skewed data, but is more sensitive to sampling variability compared to the arithmetic mean. For the above data, the median would be the third largest of the five values, which is 4.

Because the median is the middle value of a set of observations arranged in order of magnitude, it is also called the 50th percentile. In general, a *percentile* is the smallest score below which a given percentage of cases fall. For example, if a salary of £25,000 is reported as the 95th percentile, it means that 95% of the respondents have salaries *lower* than £25,000.

4.5.3 The mode

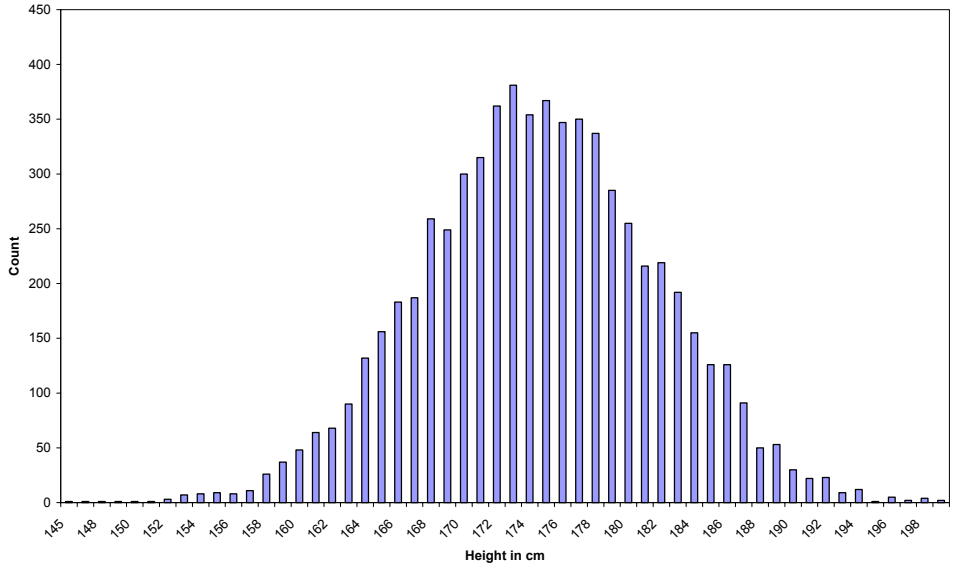
The mode is the most frequent value of a distribution. In the above example the modal value is 2.

Sometimes a distribution has more than one value with similarly large numbers of observations. This is called a *bimodal* distribution if there are two modal values or *multimodal* if there are more. Although the mode can be calculated for nominal, ordinal and interval/ratio level variables, it is the *only* measure of central tendency applicable to nominal variables. As with the median, the mode has no mathematical properties; it is also not sensitive to extreme values of the data but the most sensitive to sampling variation.

Using the example of the 6,542 men's heights collected in the HSE and referred to above, the mean value was 175cm, the median 174cm and the mode 173cm. Note that the values are similar – this is an indication that the distribution of the values is symmetric. This symmetry is shown in Figure 4.1, where the number of men for each height is approximately the same either side of the mean value (175cm). (Figure 4.1 also shows

that height is normally distributed; see Section 4.7 below for more information).

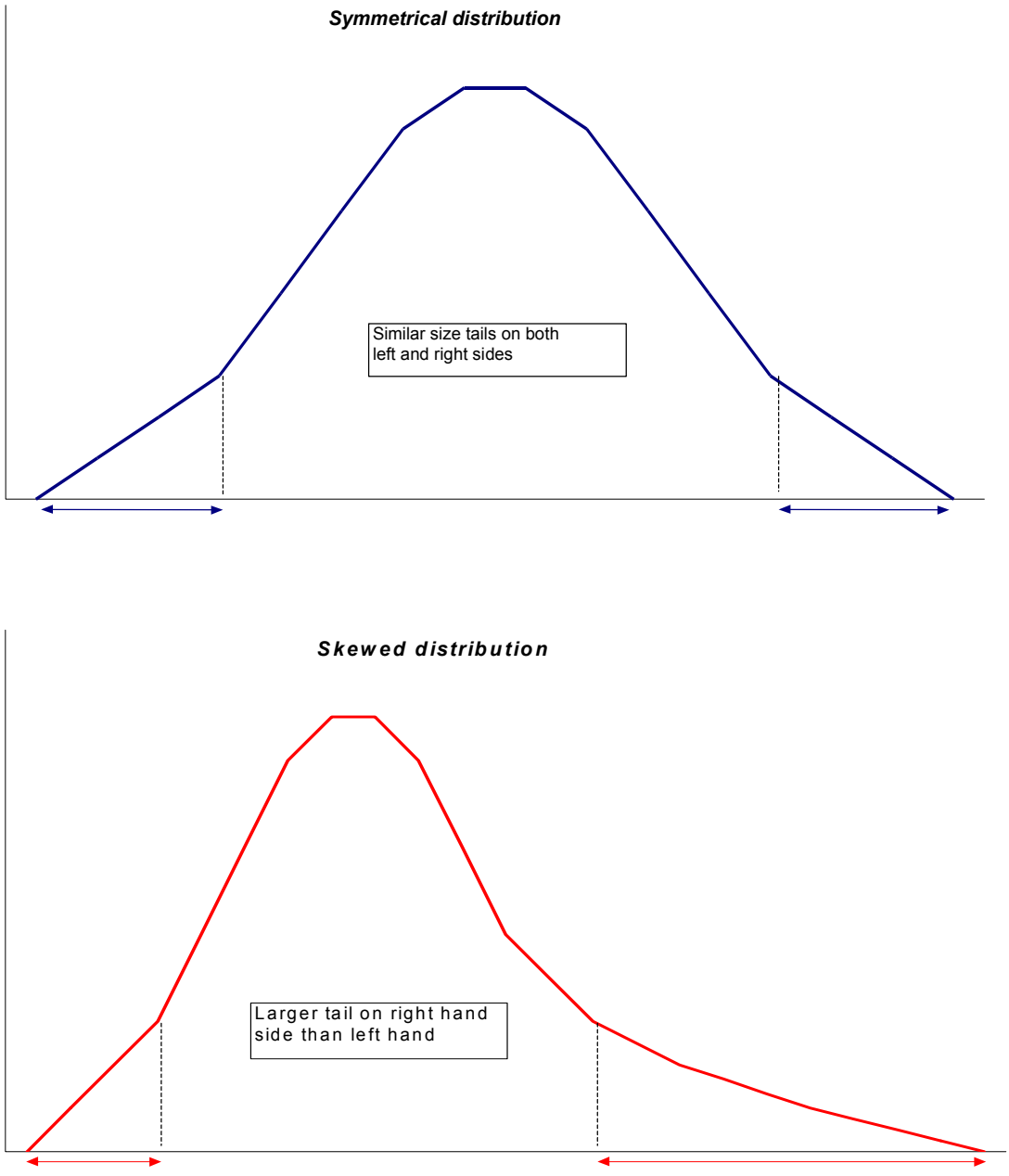
Figure 4.1 Counts of heights of men



Although the mean is the measure of average that is most often reported, when the distribution of the values is not symmetric (described as *skewed*), the median is actually a more stable measure of the average as it is not so sensitive to extreme values.

In symmetrical distributions (such as males' height described above) the two tails of the distribution are similar. With a *skewed* distribution one tail is larger than the other, depicting a few extreme values. A good example of a variable with a skewed distribution is household income – a small minority of households earn extremely large amounts! Figure 4.2 illustrates the difference between these two types of distribution graphically.

Figure 4.2 Symmetrical vs. skewed distribution

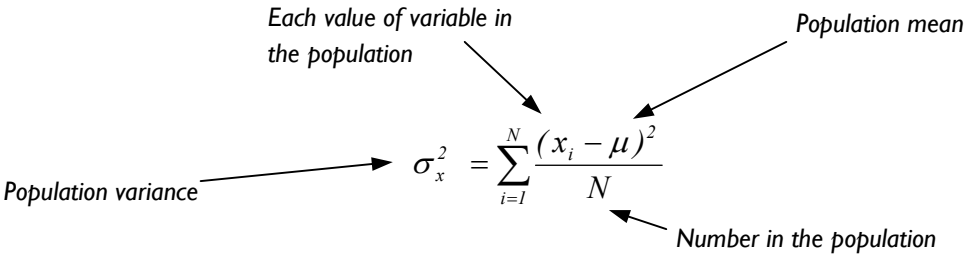


4.6 Measures of variability

In addition to estimates of the average value, a measure of the *spread* of the values is also often reported.

4.6.1 Variance and standard deviation

The measure commonly used to summarise the *spread* of data such as height is the variance, as this has the most useful statistical properties. It is defined to be the average of the squared distance of each value to the mean value.



The square root of the variance is called the standard deviation (s or SD). A value of the variance (or standard deviation) that is close to zero indicates that all the values in the sample are approximately the same and a large value implies that the observations vary widely.

4.6.2 Range

The range is the difference between the largest and smallest values and hence is likely to depend on the sample size (it will increase as the sample size increases) and be sensitive to extreme values. This is one of the weaknesses of using the range as a measure of variation.

4.6.3 Inter-quartile range

The inter-quartile range is a more stable measure of the spread, being the difference between the 25th and 75th percentile[1]. It is often used as an alternative measure of 'range' as it is unaffected by extreme values. However this measure does not share the useful statistical properties of the variance and so is less frequently used.

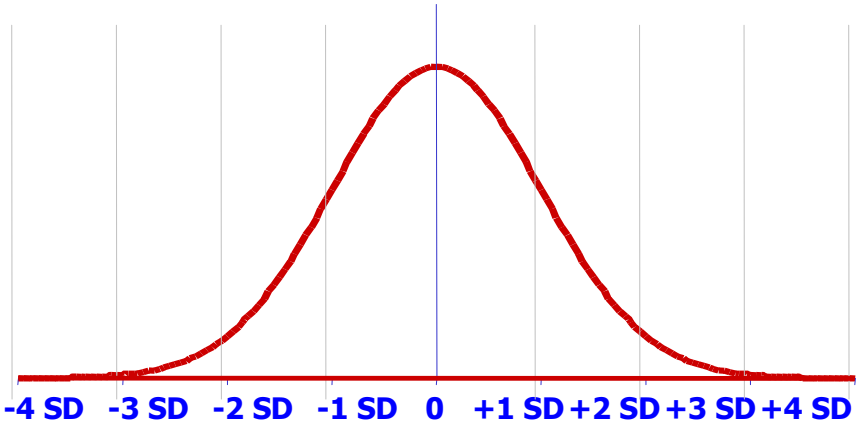
4.7 The Normal Distribution

The normal distribution is an important distribution in statistics as many 'natural' phenomena (e.g. height and weight) are normally distributed. In addition, it can be shown (using the Central Limit Theorem) that, irrespective of the distribution of the original variable, estimates of means and proportions derived from large random samples are approximately normally distributed. (In mathematical terms, the distribution of a sample estimate *tends* towards the normal distribution as the sample size increases.)

[1] Another way of expressing percentiles is the 'upper quartile' i.e. the top 25 per cent of values and the 'lower quartile' i.e. the bottom 25 percent of values.

The Normal distribution has a distinctive 'bell' shape (Figure 4.3). It is determined by two parameters, the mean (μ) and the standard deviation (σ). Once we know these values, then we know all we need about that distribution.

Figure 4.3 The standardised normal distribution



Using the properties of the normal distribution, it is possible to calculate the probability of obtaining a measure above or below any given value. It is usual to *standardise* the variable of interest to have zero mean and variance equal to one (done by subtracting the mean and then dividing by the standard deviation) so that the well-known properties of the *standard normal distribution* can be utilised. The standard normal distribution is the normal distribution with mean equal to zero and variance equal to one. Probabilities associated with this distribution are available in published tables but it is worth noting that in the standard normal distribution the area under the normal curve takes a particular form:

- 68.3% of the area is within + or – 1 standard deviation
- 95.4% of the area is within + or – 2 standard deviations
- 99.7% of the area is within + or – 3 standard deviations

4.7.1 Z-scores

There are many different normal curves each with different means and standard deviations. The *Standard Normal Distribution* has a mean of 0 and a standard deviation of 1. Standardising normal distributions to the Standard Normal Distribution facilitates comparisons.

Z-scores are a useful way of standardising variables so that they can be compared. Standardisation allows us to compare variables with different means and/or standard deviations and scores expressed in differing original units. To standardise the values of a variable, we need to take the difference between each value and the variable mean and divide each

difference by the variable's standard deviation. Statistical tables of the Standard Normal Distribution can then be used to calculate the probability of getting a smaller (or larger) z-score than the one actually obtained. Thus z-scores can also help us to identify extreme values (outliers).

For example, suppose that Claire obtained a mark of 64 in a chemistry test (where the mean was 50 and the standard deviation was 8), and Jamie obtained a mark of 74 in politics (where the mean was 58 and the standard deviation was 10). To find out who did better, we can calculate z-scores for the two marks.

Claire: $$z = \frac{x_1 - \bar{x}_1}{s_1} = \frac{64 - 50}{8} = 1.75$$

Jamie: $$z = \frac{x_2 - \bar{x}_2}{s_2} = \frac{74 - 58}{10} = 1.6$$

Thus, although Jamie's marks were higher than Claire's, Claire was 1.75 standard deviations higher than the mean, while Jamie was only 1.6 standard deviations higher. Furthermore, if we know the frequency distribution of the marks, we can calculate the percentiles for the marks. If marks for both tests had a Normal distribution, for Claire, a z-score of 1.75 corresponds to the top 4% of the class. This is obtained from the Standardised Normal Distribution tables. A z-score of 1.75 corresponds to 0.0401 of the upper tail of the Normal distribution. Jamie's mark, on the other hand, would put him in the top 5.5%.

4.8 The t-distribution

Associated with the normal distribution is the t-distribution (often called Students' t). The most important difference between the normal distribution and the t-distribution is that the distribution of the t-distribution is different depending on the sample size. The t-distribution is therefore defined by the mean, variance and the sample size (expressed as the degrees of freedom = sample size – 1). Because the t-distribution tends to the normal distribution as the sample size increases, it only makes a difference in practice when the sample size is relatively small (e.g. $n < 100$).

4.9 Confidence intervals

A confidence interval gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. In other words because the estimate is based on a sample rather than the full population, it deviates from the population values by an amount that varies according to the particular sample selected. This variation of a sample estimate from the true population value implies that it is not possible to report the exact

population value based on a sample of the population. However, through sampling theory, it is possible to state a range of values within which one is fairly sure that the true population value lies. This range is the *confidence interval.*

Because estimates of the population parameters that are derived from the sample have an approximate normal (or t-) distribution, this can be used to describe the accuracy of any estimate and, in particular, to derive confidence intervals within which we are fairly sure the *true* population value lies.

To generate the 95% confidence interval, we use the property of the standard normal distribution that the probability of a value being in the range (-1.96, 1.96) is 0.95.

The confidence interval itself is then calculated from both the survey estimate and the *standard error*.

The standard error is a variance estimate that measures the amount of uncertainty (as a result of sampling error) associated with a survey statistic. It can be thought of as a measure of the theoretical *spread* of the estimates that would be observed if a large number of samples were selected, rather than just one. For example, the standard error of the mean of a sample is:

Standard error of the sample mean

$$s.e.(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Population standard deviation

Number in the sample

As ☐ is a population parameter, it is usually not known when a sample has been selected. However, it can be estimated from the units in the sample by:

Estimate of the standard deviation

$$s_x = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

Each value of variable x in the sample

Sample mean

Number in the sample

The higher the value of the standard error, the greater the sampling variability and hence the wider the confidence interval required to be fairly sure that it includes the true population value. The value of the standard error is affected by two components:

- the amount of variability in the population of the characteristic being measured - the greater the variance in the population, the greater the variability in the survey estimate; and

- the size of the survey sample - the larger the survey sample, the more precise the survey estimate and hence the smaller the standard error.

To demonstrate the derivation of a confidence interval, we will generate a confidence interval for the number of portions of fruit and vegetables consumed by the HSE sample in a 24-hour period. The mean number of portions consumed by the sample of respondents in the HSE 2001 was 4.61, with an associated standard error of 0.02. The formula to generate a confidence interval within which we are 95 per cent certain that the population value will fall is:

CI = mean ± 1.96 x standard error

Therefore the confidence interval ranges from *4.61 – (1.96 x 0.02) = 4.57 to 4.61 + (1.96 x 0.02) = 4.64*. Hence, we are 95 per cent 'confident' that the true population value of the mean number of portions of fruit and vegetables consumed is contained within the range (4.57, 4.64).

It is important to clarify exactly what is meant by a confidence interval. As the confidence interval is estimated from a sample of the population, it has a random component due to the sampling variability[2]. The formula for the confidence interval is therefore derived to satisfy the theoretical notion that if a large number of samples were drawn, we would expect the resultant confidence intervals to include the population value 95 times out of 100. Therefore in the example above, the confidence interval for the sample selected was (4.57, 4.64). Different samples would generate different confidence intervals, of which 95 per cent would contain the population value.

4.10 Presenting data

4.10.1 Tables

One of the most common methods for describing data is the use of tables. Tables can be used to show estimates for both the full population and for sub-groups of the population - the latter allowing differences across sub-groups to be examined. Summary measures that are commonly shown in tables include means and medians, variances and inter-quartile ranges, percentages and totals.

The mean, which was introduced earlier in this chapter (see section 4.5.1 above), is a measure of the *centre* (or average value) of the distribution of a set of values. In addition to estimates of the average value, a measure of the *spread* of the values is also often reported. Measures commonly used to summarise the *spread* around the mean or

[2] Note that it is the confidence interval that is subject to random variation and not the population value, which is fixed.

median are the variance (expressed as the standard error) and the inter-quartile range respectively (see section 4.6.3 above).

The measures of average described above are useful when describing *continuous* measures, such as height or weight. However, many characteristics are measured by assigning cases to one of a number of categories, e.g. gender is collected using two categories – 'male' or 'female'. To summarise these *categorical* measures it is often not possible or indeed meaningful to generate a measure of the average – for example, a measure of the 'average gender' of the population is meaningless – and so other techniques to summarise the data are employed. One simple technique is to report the percentage (or proportion) of the population that falls into each category.

Some measures that that are collected as *continuous* variables are subsequently coded into categories so that they can be summarised using percentages. For example, age is often obtained as a continuous measure, but then categorised into age groups (e.g. 20-29, 30-39 etc.) for reporting purposes.

To demonstrate the summary measures described above and the inferences that can be made from them, we will use a table that appeared in the report for the HSE 2001 (Doyle and Hosfield, 2003). This table (Table 4.1) summarises the number of portions of fruit and vegetable consumption in a 24-hour period for different age groups. Note that the measure of fruit and vegetable consumption is collected as a continuous variable (derived from the amount of different types of fruit and vegetables consumed), from which a categorical measure (none, less than one portion, etc.) has been generated.

The first rows of the table summarise the distribution of the amount of fruit and vegetables consumed for each age group and for the population as a whole, by showing the percentages that fell into each range of the amount consumed. Various conclusions can be drawn from these percentages. For example, 7% of adults consumed no fruit, with young people in the age range 16 to 24 least likely to have consumed fruit and vegetables (13%).

The table also shows the mean and median number of portions of fruit and vegetables. The estimates for the mean show that, on average, adults in the HSE 2001 had consumed 4.61 portions of fruit and vegetables, with young people in the age range 16 to 24 consuming the least fruit and vegetables (mean=2.8). The estimates of the median show that 50% of adults consumed at least 4 portions of fruit and vegetables. Consistent with the other estimates, the value of the median is lowest (2.0 portions) for young people aged 16 to 24.

Table 4.1 Fruit and vegetable consumption, by age (adults)

Fruit and vegetable consumption (portions per day)	Age							Total
	16-24	25-34	35-44	45-54	55-64	65-74	75+	
	%	%	%	%	%	%	%	%
None	13	8	8	6	4	4	4	7
Less than 1 portion	3	2	2	3	3	3	3	3
1 portion or more but less than 2	22	19	16	15	13	14	16	16
2 portions or more but less than 3	21	19	20	16	15	15	18	18
3 portions or more but less than 4	14	16	16	16	16	19	19	17
4 portions or more but less than 5	10	13	13	14	16	14	16	13
5 portions or more but less than 6	7	8	9	11	13	11	10	10
6 portions or more but less than 7	4	5	6	7	8	8	6	6
7 portions or more but less than 8	3	4	4	5	4	5	4	4
8 portions or more	4	6	6	8	8	6	5	6
Mean	2.82	4.38	4.54	4.88	4.09	4.85	4.58	4.61
(Standard error)	*(0.06)*	*(0.05)*	*(0.05)*	*(0.06)*	*(0.07)*	*(0.06)*	*(0.06)*	*(0.02)*
Median	2.0	4.0	4.0	4.3	4.7	4.5	4.3	4.0
Base	*1774*	*2586*	*3041*	*2690*	*2210*	*1912*	*1434*	*15647*

Note that because of the distribution of fruit and vegetable consumption, the mean is not necessarily the most robust measure of the average in this example. This is because the value of the mean is disproportionately influenced by the relatively few people who consumed a large number of portions of fruit and vegetables. (This is reflected in the relatively large difference between the mean and the median – an indicator that a distribution is skewed.) For such skewed distributions, the median is a more stable measure of the average value, because it is less influenced by the extreme values.

4.10.2 Graphs

In order to more clearly and/or further emphasise particular characteristics of the data, graphical methods are often used. We shall demonstrate this using the data in Table 4.1 above.

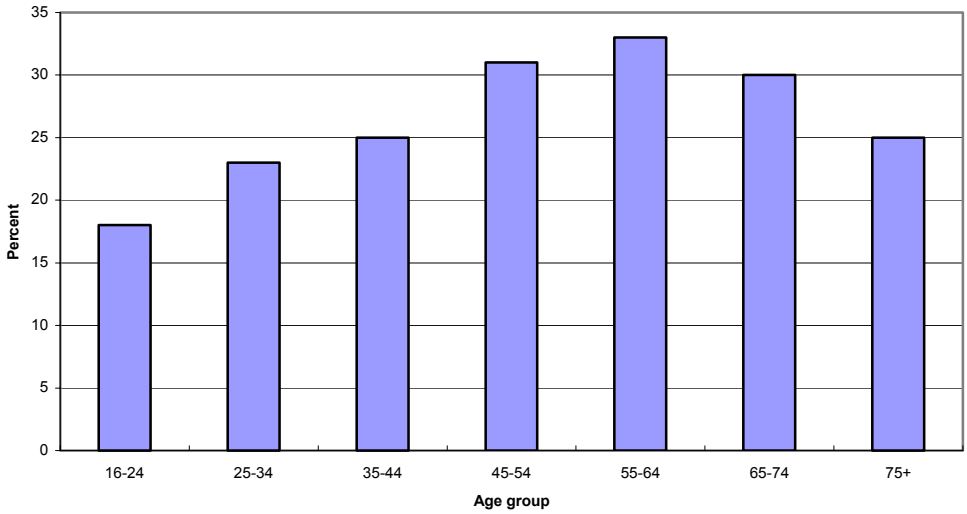
The World Health Organisation recommends that people consume at least five portions of fruit and vegetables each day (WHO, 1990). The data from the HSE 2001 can be used to examine how the proportion of people consuming at least the recommended amount of fruit and vegetables varies with people's characteristics - in this example, age. By combining rows in the table above, we can show that only about a quarter (26%) of adults consumed at least the recommended daily amount of five portions of fruit and vegetables (in the 24 hour period). To show the differences across the age group, we could combine the

rows for all the age groups and report the combined percentages. Alternatively, to give particular emphasis to the differences across the age groups, a simple bar graph can be produced (see Figure 4.4).

Figure 4.4 shows clearly that the proportion that had consumed five or more portions of fruit and vegetables increased with age up to a peak for the age group 55 to 64 after which it reduced. This relationship between fruit and vegetable consumption and age is more clearly presented by the graph, than by the same information in a table.

Various different types of graph could have been used. For instance, a pie chart, scatter plot, or a doughnut graph could have been used to represent the data. However, in Figure 4.4 a bar chart has been used.

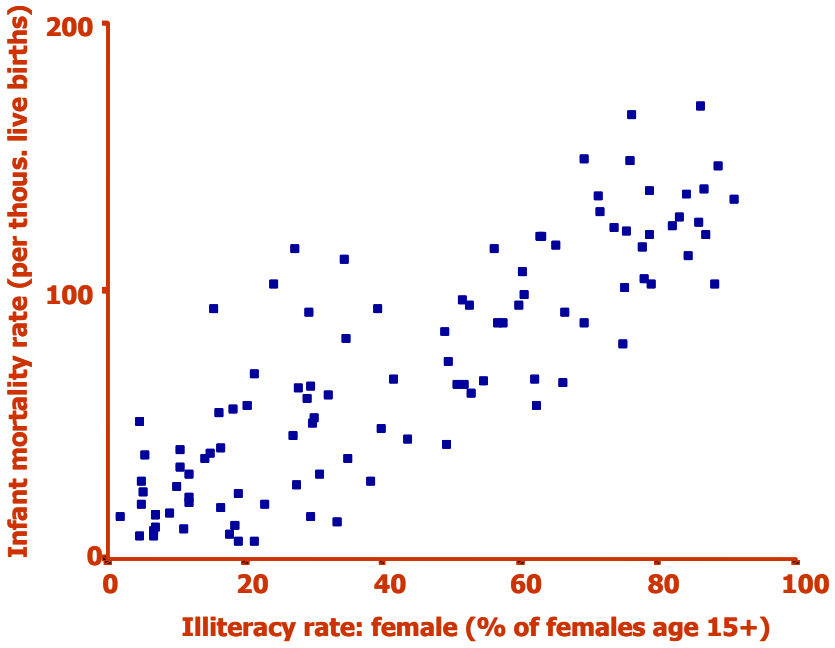
Figure 4.4 Proportion of adults consuming five or more portions per day, by age group



4.10.3 Graphs - Scatterplots

One way of visualising the relationship between two continuous variables is to create a scatterplot. Scatterplots provide graphical tools for exploring the distributions and relationships of the two continuous variables. The question we try to answer is whether or not the variation in one variable (dependent variable Y) can be explained by the variation of the other (independent variable X). For example Figure 4.5 shows the scatterplot for female illiteracy rate by infant mortality (source: World Bank, 1992):

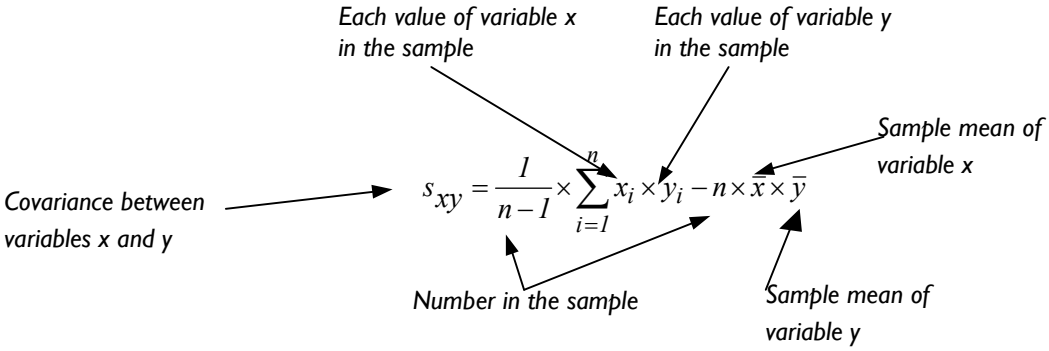
Figure 4.5 Scatterplot of female illiteracy rate by infant mortality



4.11 Relationships between pairs of variables

4.11.1 Covariance

Scatterplots are useful tools for exploring the data by providing a visual presentation of the relationship between X and Y, but do not provide a measure of the strength of the relationship. An appropriate numerical measure (i.e. statistic) is needed in order to draw any conclusions about relationships between variables. *Covariance* (s_{xy}), is a measure that reflects the strength of the (linear) association between two variables. It is defined as:



For example, the following data (Table 4.2, taken from 'Basic statistics for behavioural sciences' by Heiman, G. W., 1996, Houghton Mifflin, 2nd edition) can be used to ask the question whether the amount of coffee one drinks is related to a person's degree of nervousness?

Table 4.2 Covariance example

No. of cups of coffee (X)	Nervousness score (Y)	Product (XY)
1	1	1
1	1	1
1	2	2
2	2	4
2	3	6
3	4	12
3	5	15
4	5	20
4	6	24
5	8	40
5	9	45
6	9	54
6	10	60

$$\bar{x} = 3.308 \qquad \bar{y} = 5.000 \qquad \sum x_i \times y_i = 284$$

$$\textit{Therefore} \qquad s_{xy} = \frac{1}{12} \times (284 - 13 \times 3.308 \times 5.000) = 5.750$$

Ideally, we want to detect a large value of covariance, which would indicate stronger linear relationship between X and Y. However, the problem with the covariance is that it is dependent on the scale of measurement of either variable. If nervousness score has been measured out of 100 rather than out of 10 we would have arrived at a value for the covariance, which was 10 times as great. Consequently, one can make the covariance between X and Y as large (or as small) as one pleases, without changing in any way the 'shape' of the joint distribution of the two variables.

4.11.2 The correlation coefficient

In order to get round the problem of differing units, and to get a measure that can be used to compare between different pairs of variables, we would need to standardise our measure. *Correlation* is a measure of association derived from standardised variables, or in other words, a standardised covariance. Each variable is standardised by subtracting the mean and dividing by the standard deviation[3].

A common misconception is that high correlation is equivalent to *causation*. That is, if two variables X and Y have a correlation close to 1, then we can assume that either X causes Y or vice versa. This assumption is *incorrect.* Correlation *does not* prove causation. A large correlation between X and Y *does not* mean X causes Y (or vice versa). It is just a mathematical measure of the strength of the relationship

[3] Note that standardised variables have a mean of 0 and a standard deviation of 1.

between the two variables. A high correlation between X and Y could be because:

- X and Y influence one another.
- X and Y co-vary; they exist together as part of a single system. For example, heartbeat and breathing exist together and are part of the same system, but neither causes the other in any illuminating sense of the word.
- the relationship between X and Y is actually due to other variable(s) (Zs) which influence both X and Y. For example, one might observe a strong correlation between the number of violent crimes and the number of religious meetings in a sample of cities. This correlation does not mean that religious meetings lead to violent crimes. It is most probably due to the fact that both variables are proportionately related to the population of the city, in that the number of violent crimes and the number of religious meetings tend to increase linearly with the size of the city. Consequently, we have a high correlation between city size and violent crime, as well as between city size and number of religious meetings, resulting in a high correlation between incidents of violent crime and number of religious meetings.

Another misconception is that a low correlation coefficient suggests that the relationship between X and Y is weak or low. This interpretation is true only for a *linear* association. It is possible for two variables to have a very strong relationship that is non-linear but the correlation coefficient would not be able to pick this up. For example, there is a strong relationship between an offender's age and their likelihood of reconviction but the correlation coefficient indicates a weak relationship. This is because the relationship between age and reconviction rates is non-linear. Offenders are more likely to reconvict at different stages of their life cycle, with rates increasing up to early adulthood but then declining in later years. Non-linear relationships can be detected by the use of other statistical methods. For example, ages could be grouped into bands and each band entered into a regression model as a categorical variable (see Table 4.4 for an example).

A correlation, whether linear or non-linear, can infer a causal relationship between two variables if:

- *The relationship is plausible*. Many statistical relationships are coincidental. For example, there was a high correlation between the proportion of marriages taking place in church and the death rate during the years 1866 to 1911. That is, in the years when people were most likely to get married in church the death rate was higher than in years with a smaller proportion of church weddings (Rowntree, 1981). The relationship though was the result of chance. There was no

logical reason to suppose that getting married in church caused people to die. In order to infer a causal relationship the correlation therefore needs to be logical.

- *Cause precedes effect*. A change in the explanatory variable should be followed by a change in the dependent variable. For example, a treatment could only be said to be effective if the patient's symptoms improved after he/she had received the treatment.

- *Alternative explanations are eliminated*. The surest way to eliminate confounding explanations would be to conduct a randomised controlled trial. This method would control for extraneous variables that might confound the results. Having completed the randomised controlled trial, if one can establish that the experimentally manipulated variable is correlated with the dependent variable, then one can make a causal inference.

Finally, we mentioned that the correlation coefficient depends on several assumptions (i.e. continuous scale, linearity and normality). When these assumptions are unreasonable, we should use other measures of association that have less stringent assumption, such as:

- Spearman's rank correlation coefficient (ordinal data);
- Kendaull's Tau (ordinal data);
- Phi coeficient (for a 2 x 2 table and categorical data);
- Odds ratio (for a 2 x 2 table and categorical data).

4.12 Regression

4.12.1 Simple linear regression

As said, correlation does not imply causation or the ability to 'explain' a dependency of one variable on another. It is simply a measure of association that tells us whether two *continuous* variables vary together. If we are interested in trying to 'explain' the behaviour of one variable (the dependent variable) using the predictive power of another variable (the independent or predictor variable), we need to use *simple regression analysis*. If we have two or more independent variables we would use *multiple regression analysis*.

Consider the scatterplot (Figure 4.6) of the percentage of the adult population (16+) in a Health Authority (HA) with limiting long-standing illness by the percentage of households in the HA in non-manual socio-economic group (SEG 1-6 and 13) according to the 1991 Census of the population.

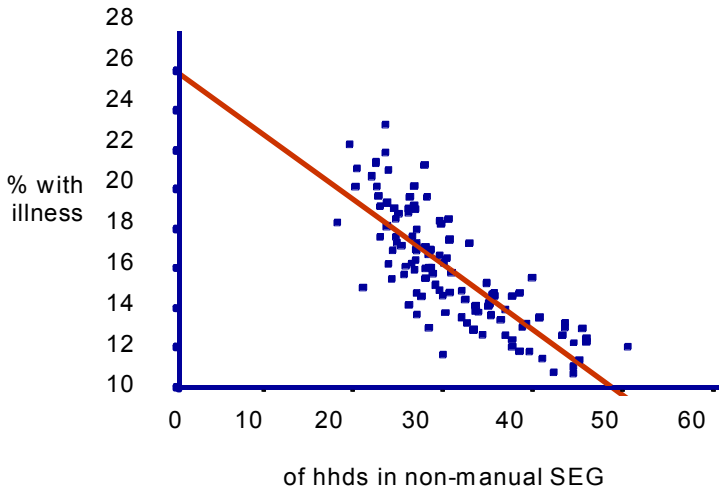
Figure 4.6 Scatterplot of limiting long-standing illness by non-manual SEG

Here, long-tem illness (which is usually taken as a measure of good health) is the dependent variable, and non-manual SEG (which is usually taken as an economic measure) is an independent or predictor variable. The assumption (model) is that economic conditions as measured by the household SEG profile for a HA may predict, or explain, the prevalence of long-term illness in the HA.

As we can see, there is a downward trend in the data points, which is an indication that as the percentage of non-manual households in the area (HA) increases, the incidence of long-term illness decreases. However, we can now go beyond simple correlation where the variation in the prevalence of long-term illness is reflected by a similar variation in the socio-economic profile of the households in an area and try to test whether we can *explain* this variation. That is, we can test the hypothesis that high non-manual household rates *lead* to lower illness rates in the area.

It looks like most points lie close to a straight line (see Figure 4.7). In other words, there is an approximately (negative) linear relationship between non-manual household rates and long-term illness.

Figure 4.7 Scatterplot with fitted line



4.12.2 Simple linear regression – the equation of a straight line

The equation of a straight line (in the population) can be written as:

$$Y = \alpha + \beta \times X$$

where (see also Figure 4.8),

– Y is the *continuous* dependent variable.

– X is the *continuous* independent variable.

– a is the intercept, which is the value of the dependent variable for $X = 0$ (or the value where the line crosses the Y-axis).

– β is the slope which describes how much the value of the dependent variable changes when the independent variable increases by 1 unit (for the above example 1 unit = 1%).

However, using sample data (as we saw in Figure 4.7), not all data points will fall on the line. Some data points (*observed* values of Y - y_i) will be very close to the line while others will be more distant. Therefore, the fitted line through the data points is only an approximation. In this case, the equation of the straight line can be used to *predict* the dependent variable. The *predicted* values of Y ($\hat{y}_i$), are located on the line.

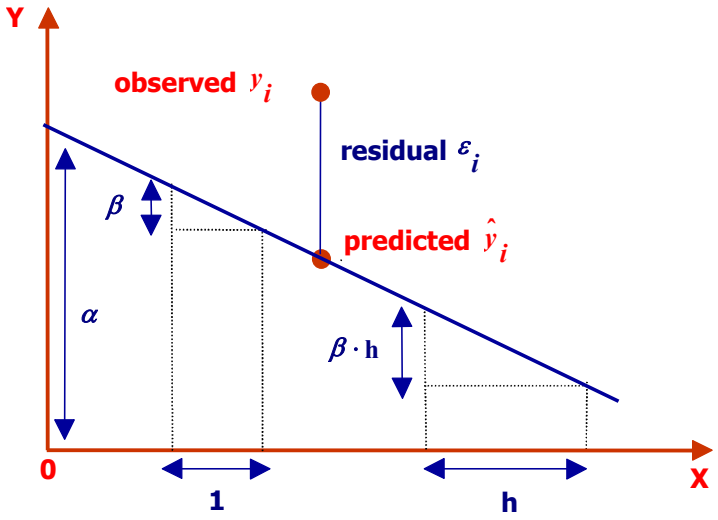
The difference between the observed and the predicted value is called the *residual.* This represents the difference between a particular data point and the line. The sum of the squares of all residuals (or 'deviations') is called the *residual* or *error sum of squares* (SSE or SS_{RES} for short).

Therefore, we can write:

$$\hat{Y} = \alpha + \beta \times X + \varepsilon$$

which says that a value of the dependent variable can be predicted by a value of the independent variable multiplied by a coefficient (the slope) plus a constant factor (the intercept) plus an error term (the residual). This error term represents the effects of all the other factors that have an influence on the dependent variable other than the independent variable in the equation.

Figure 4.8 The regression line



The best fit to the data is the line that has the minimal SSE, that is the line that minimises the distances between the data points and the line. The method used to position the line in such a way is called *Ordinary Least Squares* (*OLS*) and the line that best fits the data is the *OLS regression line.* The OLS regression line is useful for forecasting and prediction (by assuming that *all* data points are on the regression line). An example of regression analysis is presented at Box 4.1 below.

4.12.3 Multiple regression

Multiple regression is an extension to simple regression to include two or more independent (explanatory) variables.

Consider the Health Authority (HA) example, where economic conditions as measured by the household SEG profile for a HA was used to predict (or explain) the prevalence of long-term illness in the HA. It is likely that there are many other social and economic factors that may affect the incidence of long-term illness in an area other than the socio-economic profile of the households in the area. Such factors may include unemployment rate, population density, the access and use of a car, educational qualifications etc. For example, we might suppose that the prevalence of limited long-term illness would be higher in areas with high unemployment rate than in areas with low rates irrespective of the socio-economic profile of the area.

Whether or not other factors have an effect on the dependent variable can be investigated by adding further independent variables to the model.

In general, the multiple regression model *in the population* can be expressed as:

$$Y = \alpha + \beta_1 \times X_1 + \beta_2 \times X_2 + ... + \beta_p \times X_p$$

where,

- Y is the *continuous* dependent variable.
- X_p is the *continuous* independent variable.
- p is the number of independent variables in the model.
- α is the *regression coefficient* for the intercept
- β_p is the regression coefficient associated with the independent variable X_p.

In the same way as in the simple regression model, there is an underlying assumption of a linear model in the population with the observed values of the dependent variable (y_i) being statistically independent of each other. Also it is assumed that there is a normal distribution (multivariate normal distribution) of the dependent variable for every combination of the values of the independent variables in the model with constant variance. For example, if illness rates is the dependent variable and household non-manual rate and unemployment rate are the independent variables, it is assumed that there is a linear relationship between the dependent and independent variables. Also, for every combination of non-manual rate and unemployment rate there is a normal distribution of illness rates, and though the means of these distributions may differ, all have the same variance.

The regression coefficients α and $\beta_1, \beta_2 ... \beta_p$ are unknown population parameters and are estimated from the sample by $\hat{\alpha}$ and $\hat{\beta}_1, \hat{\beta}_2 ... \hat{\beta}_p$ respectively. Therefore, the *estimated* regression model is:

$$y_i = \hat{\alpha} + \hat{\beta}_1 \times x_{1,i} + \hat{\beta}_2 \times x_{2,i} + ... + \hat{\beta}_p \times x_{p,i} + \varepsilon_i$$

where,

- y_i is the value of the *continuous* dependent variable Y for case *i*
- $x_{p,i}$ is the value of the *continuous* independent variable X_p ,for case *i*
- $\hat{\alpha}$ is the *estimated* regression coefficient for the intercept
- $\hat{\beta}_p$ is the *estimated* regression coefficient associated with the independent variable X_p .

– ε_i is the residual, that is the difference between the observed and predicted (from the regression line) value ($\varepsilon_i = y_i - \hat{y}_i$).

Estimation of $\hat{\alpha}$ and $\hat{\beta}_1, \hat{\beta}_2 \ldots \hat{\beta}_p$, is via ordinary least squares (OLS) in the same way as in simple regression: we choose $\hat{\alpha}$ and $\hat{\beta}_1, \hat{\beta}_2 \ldots \hat{\beta}_p$ that minimise the error sum of squares $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. The solution is very messy when written using sums but is much better using matrix notation. For the formulae for $\hat{\alpha}$ and $\hat{\beta}_1, \hat{\beta}_2 \ldots \hat{\beta}_p$ see, for example, Draper and Smith (1998) *Applied Regression Analysis*, Third Edition, Wiley.

Each regression coefficient $\hat{\beta}_1, \hat{\beta}_2 \ldots \hat{\beta}_p$ reveals the amount by which the dependent variable is expected to change with a 1-unit change in the independent variable when the other independent variables are held constant (controlled for). Or, in other words, the regression coefficients $\hat{\beta}_1, \hat{\beta}_2 \ldots \hat{\beta}_p$ provide information on the impact of each independent variable on the dependent variable while *simultaneously controlling* for the effects of the other independent variables in the model.

Box 4.1 How to interpret the regression equation: a hypothetical example

Imagine we have run a simple linear regression analysis on house price data. We are interested in the effect of investing money (measured in £000s) in home improvements on property values. The value of the property (measured in £000s) is therefore the dependent variable and the value of home improvements is the independent variable. In other words we are using regression analysis to predict property values from the value of home improvements.

The results of the analysis gives us α =150 and β=1.5. β indicates the average change in the dependant measure, corresponding to one-unit change in the independent variable. The regression formula would be…

Predicted property value = 150 + (1.5 x value of home improvements)

This indicates that on average a £1,000 investment in home improvements is accompanied by a £1,500 increase in house value. The intercept (α) suggests that if there was no investment in the house, we would expect it to be worth £150,000.

However, we can hypothesise that house value depends on more than just the amount spent on home improvements. To test this hypothesis we could add more independent variables, such as the number of bedrooms (β_2) and conduct a multiple regression. This time the analysis gives us α =130, β_1=1.2 and β_2=50, which translates into the following equation…

Predicted house value = 130 + (1.2 x value of home improvements) + (50 x number of bedrooms)

From this we can calculate the expected house value. For example we would expect a studio flat with no investment in home improvements to be worth on average £150,000 [150 + (1.2x0) + (50x0)]. Whereas we would expect a five-bedroom home where the owner(s) have invested £50,000 to be worth £460,000 [150 + (1.2x50 + 50x5)].

4.12.4 Logistic regression

Another one of the most common multivariate techniques is logistic regression used when the dependent variable is binary (i.e. only has two possible outcomes). To illustrate the use of logistic regression, we will use a hypothetical example looking at the relationship between pet ownership and limiting long-term illness. First, we will look at the simple bivariate relationship between the two variables. Table 4.3 shows a

simple two-way table of having a limiting longstanding illness against pet ownership. This shows that the prevalence of limiting longstanding illness is lower amongst pet owners (24.6%) than non-pet owners (27.8%) by over three percentage points – a difference that is highly statistically significant[4] [5].

Table 4.3 Bivariate relationship between limiting longstanding illness and pet ownership

	No limiting long-term illness	Limiting long-term illness
No pet	72.2% *6,207*	27.8% *2,387*
Pet	75.4% *5,306*	24.6% *1,734*

Therefore it appears from looking at the bivariate relationship that there is a link between pet ownership and limiting longstanding illness and that the rate of limiting longstanding illnesses is lower (albeit not by very much) amongst pet owners. The naive researcher might even infer from this result that owning a pet reduces the risk of having a limiting longstanding illness. However, the link between limiting longstanding illness and pet ownership is not as simple as this bivariate relationship suggests. In order to examine further the link between the two variables, we will fit a logistic regression model, which will allow us to control for other characteristics.

Table 4.4 shows the results of fitting a logistic regression model of limiting longstanding illness on pet ownership, age and sex. The estimates of the odds ratios show the multiplicative increases (or decreases) in the odds of having a limiting longstanding illness for each category compared to the baseline category. So for example, the odds for women of having a limiting longstanding illness in the model are greater than the odds for men by a factor of 1.140. Controlling for sex and age in the analysis gives a completely different interpretation of the relationship between limiting longstanding illness and pet ownership. From the logistic regression model, there is evidence that owning a pet is associated with a slight increase, rather than decrease, in the rate of limiting longstanding illness (although this effect is very small and only marginally significant). This is signified by the value of the odds ratio being greater than 1 (1.086) for pet ownership, showing that the odds of having a limiting longstanding illness is increased (by a factor of 1.086) for

[4] For a discussion of statistical significance see section 4.13.1

[5] $\chi^2_1 = 19.71, p < 0.001$

pet owners. The increase is described as marginal because the associated p-value of 0.04 is only slightly smaller than the 0.05 significance level.

Table 4.4 Logistic regression of limiting longstanding illness on pet ownership, age and sex

Covariate	odds ratio	s.e.(odds)	z-score	p>|z|	95% Confidence Interval	
Own pet:						
No pet	1.000	(baseline)				
Pet	1.086	0.043	2.09	0.04	1.005	1.174
Age group:						
16 to 24	1.000	(baseline)				
25 to 34	1.146	0.108	1.45	0.15	0.953	1.379
35 to 44	1.615	0.142	5.44	<0.01	1.359	1.919
45 to 54	2.705	0.233	11.55	<0.01	2.285	4.202
55 to 64	4.290	0.371	16.85	<0.01	4.621	5.082
65 to 74	5.411	0.476	19.20	<0.01	4.554	6.428
> 75	8.672	0.797	24.51	<0.01	7.243	10.384
Sex:						
Men	1.000	(baseline)				
Female	1.140	0.044	4.42	<0.01	1.058	1.230

The reason why there was a spurious relationship in the bivariate relationship between limiting longstanding illness and pet ownership is *that both these variables are related to age*. Table 4.4 also shows the rates of pet ownership within each age group. As can be seen, the age profile of people that owned pets is very different to those that did not, with relatively more people aged 65 or more in the group of non-pet owners. Therefore the apparent higher rate of limiting longstanding illness amongst people who did not own pets was observed because they were more likely to be *older* than those that owned pets. Once age was controlled for (along with sex), the relationship between the two variables, allowing for age and sex differences, was revealed to be quite different.

It is worth noting that even after correcting for characteristics of the two groups, one still has to take care when interpreting the results. In the example above, it would still not be correct to state that *owning a pet slightly increases the likelihood of getting a limiting longstanding illness.* This statement suggests that there is a causal link between pet ownership and illness which has not been shown by the analysis. The analysis has merely shown that there is a statistical relationship between the two variables. There could be any number of competing explanations as to why this relationship has been observed. It is quite possible people

are more likely to get a pet *after* suffering a limiting longstanding illness, for example[6].

4.13 Statistical hypotheses

4.13.1 Hypothesis testing

Often the difference between two estimates, or between an estimate and a specific fixed value, will be reported as *statistically significant*. This implies that the difference is large enough that it is unlikely to have been observed simply because of sampling error (or, put the other way, that it is likely to have been observed because of a real difference in the population). The test to determine whether a difference is significant or not involves (often implicitly) the notion of a *hypothesis test*. In statistical terms, a hypothesis test is undertaken to ascertain if there is enough evidence to reject one hypothesis about the population (the *null hypothesis*) in favour of another (the *alternative hypothesis*) using estimates from the sample.

In most cases, the null hypothesis is the 'default' state – e.g. that a value is zero or that the difference between two values is zero (although there are exceptions to this). The alternative hypothesis tends to be the opposite of the null hypothesis – e.g. that a value is not zero or that the difference between two values is not zero.

For the purposes of the hypothesis test, it is assumed that the null hypothesis is true. With that assumption, the likelihood of an equal or more extreme value than that measured being observed is calculated. Only if it is found that it is unlikely that the measured (or more extreme) value could have been observed if the null hypothesis is true, is the null hypothesis rejected in favour of the alternative hypothesis[7].

Associated with the hypothesis test is the *level of significance*. The level of significance is the threshold that is used to decide if an observed difference in the sample was *unlikely* to have been observed by chance and hence to reject the null hypothesis. The level of significance is expressed as a probability and is often taken to be 0.05. This may also be described as significant at the 5% or 95% level, or displayed as $p<0.05$. A significance level of 0.05 implies that a difference extreme enough to reject the null hypothesis by chance when the null hypothesis is actually true will be observed one time in twenty. In some circumstances a value less than 0.05 (usually 0.01) might be used as a threshold to determine statistical significance, if one needs to more certain that the observation was not the result of chance.

[6] To fully investigate this link, it would be necessary (at the very least) to collect histories of pet ownership and illness.

[7] A helpful analogy to hypothesis testing is that of British Law. A defendant is assumed to be innocent (the null hypothesis) unless there is sufficient evidence that his or her innocence is highly unlikely, in which case the defendant is declared to be guilty (the alternative hypothesis).

To undertake the hypothesis test, the probability of obtaining a value more extreme than that observed is estimated, and that probability compared to the significance level. If the probability is smaller than the significance level, i.e. the difference is more unlikely to have been observed than the threshold level, then the null hypothesis is rejected in favour of the alternative hypothesis.

As a simple example, consider a test of whether men and women are the same height. The null hypothesis would be that men and women are (on average) the same height and the alternative hypothesis that men and women are not (on average) the same height. Although these hypotheses relate to the population, they can be tested using survey data with the appropriate allowance for the sampling variability.

The HSE 2001 recorded the average height for men as 174.52cm (s.e. = 0.09cm) and for women as 161.02cm (s.e. = 0.08cm), based on sample sizes of 6,542 and 8,011 respectively. The difference is therefore 14.50cm and, using the formula in Appendix G1, we can calculate the standard error of this difference (s.e. = 0.12cm). In order to undertake the hypothesis test, we assume that the null hypothesis is true – namely that there is no difference in height between men and women. To reject the null hypothesis (that the difference in the population is zero) and accept the alternative hypothesis (that the difference in the population is not zero), we have to test whether the difference of 14.50cm is far enough away from zero for it to be unlikely that it was observed by chance.

In fact, in this example it is found that the chance of observing a difference of 14.50cm by chance is very small[8]. Therefore there is sufficient evidence to reject the null hypothesis at the 0.05 significance level in favour of the alternative hypothesis that men and women are not (on average) the same height.

4.13.2 Hypothesis testing for categorical data

When the analysis variables are categorical, hypothesis testing can be used to compare two proportions (or percentages) or measure the association between the variables. Suppose we were interested in whether or not there was a relationship between gender and ownership of a car (using the health survey data). There are a number of ways to check this:

1. Calculate a confidence interval for the difference in the proportion owning a car between men and women. If the confidence interval does not include 0 then there is a significant difference in the proportions

2. Use the t-test for proportions

[8] This is estimated using a t-test for the difference between two means for independent samples.

3. Use the chi-square test[9]

4.13.3 The chi-square test

The rationale behind the χ^2 test is that if the two variables are not related (i.e. gender is not related to car ownership) then we should have the same proportion of men and women owning a car. There may be a difference in the proportions due to pure chance, but (depending on the sample size) this difference must be small. Consequently, if the difference between the two proportions is very large, then we would be led to conclude that there is some association between gender and car ownership.

Although the test is about proportions, the actual test statistic compares the observed and expected frequencies. Essentially the chi-square test statistic compares the observed data in each cell with what we would expect to get if the null hypothesis of no difference was true, i.e. if the proportions owning a car among men and women were the same. Suppose that the observed frequencies for a 2x2 table were:

	A	B	Total
I	n_1p_1	n_2p_2	np
II	$n_1(1-p_1)$	$n_2(1-p_2)$	n(1-p)
Total	N_1	n_2	n

If the null hypothesis of no difference is true, then we would expect $p_1=p_2=p$, so that the expected frequencies will be:

	A	B	Total
I	n_1p_1	n_2p_2	np
II	$n_1(1-p_1)$	$n_2(1-p_2)$	n(1-p)
Total	N_1	n_2	n

Another way of calculating the expected frequencies is by the formula *"column total x row total divided by grand total"*.

	A	B	Total
I	$\frac{np \times n_1}{n}$	$\frac{np \times n_2}{n}$	np
II	$\frac{n(1-p) \times n_1}{n}$	$\frac{n(1-p) \times n_2}{n}$	n(1-p)
Total	N_1	n_2	n

[9] The chi-square test can also be applied to tables larger than 2X2. For larger tables, however, the interpretation is somewhat different.

The test statistic is: $\chi^2 = \sum_{cells} \frac{(Observed - expected)^2}{expected}$

When we have obtained the test statistic, we compare it with the critical value on 1 degree of freedom (2x2 table) taken from the table of the chi-square distribution. If it is greater than the critical value, we reject the null hypothesis. Alternatively, if the associated p-value is less than 0.05 (at the 95% level) then we reject the null hypothesis and accept the alternative.

Degrees of freedom (df) is the number of cells in a table that we need to know in order to know them all, given the marginal totals. The general formula for obtaining the degrees of freedom is (r-1) x (c-1), where r is the number of row categories and c is the number of column categories. This test can be illustrated using the Health Survey data. The table below (Table 4.5) shows car ownership by gender.

Table 4.5 Car ownership by gender

		Male	Female	Total
Own car	Count	1,944	1,951	3,895
	Expected count	1,833.6	2061.4	3,895.0
	% within gender	83.3%	74.3%	78.5%
No car	Count	391	674	1,065
	Expected count	501.4	563.6	1,065.0
	% within gender	16.7%	25.7%	21.5%
Total	Count	2,335	2,625	4,960
	Expected count	2335.0	2,625.0	4,960.0
	% within gender	100.0%	100.0%	100.0%

The *Null hypothesis* (H_0) is that there is no difference between men and women in the proportions owning a car. Therefore the *alternative hypothesis* (H_1) is that the proportion of men owning a car is significantly different from that of women owning a car.

The test statistic is:

$$\chi^2 = \frac{(391-501.4)^2}{501.4} + \frac{(674-563.6)^2}{563.6} + \frac{(1944-1833.6)^2}{1833.6} + \frac{(1951-2061.4)^2}{2061.4} = 58.46$$

This is larger than 4.84 (which is the critical value for 1 degree of freedom at the 95% significance level taken from the tables of the χ^2 distribution). So, we reject the null hypothesis and conclude that there is a significant difference between men and women in terms of car ownership.

4.13.4 The odds ratio

Rather than compare the difference between two proportions it can sometimes be useful to compare the odds. If p_1 is the proportion of men

owning a car and p_2 is the corresponding proportion for women, then the odds of owning a car for men is given by $\frac{p_1}{1-p_1}$ while that for women is $\frac{p_2}{1-p_2}$. The *odds ratio* (θ) measures the association between owning a car and gender. If the same proportions of men and women own cars, then the odds ratio will equal 1.

In the example above, the odds of having a car for a man is 0.833/0.167=4.99 and the odds of having a car for a woman is 0.743/0.257=2.89. The odds ratio is 4.99/2.89=1.72 (or 2.89/4.99=0.58 – either way up will do as long as the group whose odds are in the numerator is clearly stated). The odds ratio does not change if rows become columns and columns become rows. The interpretation is as follows: the odds of a man owning a car are 1.72 times the odds of a woman owning a car. That is the odds for a man owning a car are about 72% higher than those of a woman.

Note that the odds ratio is measured on the logarithmic scale and therefore, it is *wrong* to say that the odds of a woman owning a car are 72% lower. Instead, using the odds ratio of 2.89/4.99=0.58, we can say that the odds of a woman owning a car are 42% lower than those of a man.

4.14 References

Bajekal, M, Primatesta, P & Prior, G (2003) *Health Survey for England 2001*. London, The Stationery Office.

Doyle, M & Hosfield, N (2003) *Health Survey for England 2001: Fruit and vegetable consumption.* London, The Stationery Office.

Rowntree, D (1981) *Statistics without tears: a primer for non-mathematicians*. Penguin Books.

World Health Organisation (1990) Diet, Nutrition and the Prevention of Chronic Diseases. Technical Report Series: 797, World Health Organisation.

4.15 Further reading

4.15.1 Introduction to statistical methods

Gonich, L. & Smith, W. (1993) *The Cartoon Guide to Statistics*. Harper Collins, ISBN 0-06-273102-5.

Clarke, G. M. & Cooke, D. (1998) *A Basic course in statistics*, 4th edition. Arnold, ISBN 0340 719958.

Heiman, G. W. (1992) *Basic Statistics for the Behavioural Sciences*. Houghton Mifflin Company, ISBN 0-395-51546-7.

Bryman, A. & Cramer, D. (1990) *Quantitative Data Analysis for Social Scientists*. Routledge, ISBN 0-415-02665-2.

Rose, D. & Sullivan, O. (1996) *Introductory Data Analysis for Social Scientists*. Open University Press, 2nd Edition, ISBN 0-335-19617-9.

Diamond, I. & Jefferies, J. (1999) Introduction to Quantitative Methods. 6th Edition.

Fielding, J. & Gilbert, N. (2000) *Understanding Social Statistics*. Sage, ISBN 0 80397982 7.

Rowntree, D (1981) *Statistics without tears: a primer for non-mathematicians*. Penguin Books.

4.15.2 Advanced statistical methods

Pagano, R. R. (1990) *Understanding Statistics in the Behavioural Sciences*. West Publishing Company, 3rd Edition, ISBN 0-314-66792-x.

Rose, D. & Sullivan, O. (1996) *Introductory Data Analysis for Social Scientists*. Open University Press, 2nd Edition, ISBN 0-335-19617-9.

De Vaus, D. (2002) *Analysing Social Science Data*. Sage, ISBN 0 7619 5937 8

Neter, J., Wasserman, W., & Kutner, M. (1985) *Applied Statistical Models*. Richard D Irwin, Inc.

Kleinbaum, D., Kupper, L.L. & Muller, K.E. (1988) *Applied Regression Analysis and Other Multivariate Methods*. 2nd Edition. PWS-Kent Publications.

Draper, N.R. & Smith, H. (1981) *Applied Regression Analysis*, 2nd Edition. Wiley.

Agresti, A. (1996) An introduction to categorical data analysis. Wiley.

The Magenta Book: guidance Notes for Policy Evaluation and Analysis

Background paper 5: what is sampling?

Published March 2004
Updated October 2007

Government Social Research Unit

HM Treasury
1 Horse Guards Road
London SW1A 2HQ

5 What is sampling?

5.1 Introduction

In a social researcher's ideal world all data would be collected by census. Then, for example, if it proved important to know what percentage of the population had experienced a crime in the last year, or what percentage of the population smoked, all members of the population would be 'approached' (either by an interviewer, by post, or by telephone) and asked to provide the information required.

Good as this may sound, it does not happen in practice (the single exception being in the decennial population census). The reasons for not taking such a bold approach include cost (the UK census is an extremely expensive exercise), practical problems, including the fact that the general population would get very tired of answering all the questions put to them, and, perhaps most importantly, because by using sampling methods it is usually possible to get close to the estimates we need by approaching only a very small percentage of the population. This means that we collect the data we need, but the burden of providing that information is spread thinly across the population.

This chapter describes how sampling works. We start with the basics, but in later sections move on to some of the complications that arise in real-world situations. By and large we concentrate on surveys of the general population, but some reference is made to surveys of other populations such as businesses.

5.1.1 How and why sampling works

General population survey sampling relies on the notion that it is not necessary to collect data from *all* people in order to generate statistics about that population. Intuitively this seems correct. For instance, if we wish to know the average systolic blood pressure of adults in GB, then it seems reasonable that if we measure the blood pressure of a random 1 in 1000 adults then the average we get will be pretty close, if not exactly the same, as the average we would get if we measured all adults. Or, putting the problem another way, it seems reasonable to assume that if we measure the blood pressure of, say, 10,000 adults selected at random and then take the average, then repeating the exercise with 10,000 more adults and adding the two samples together, is unlikely to generate a very different estimate. After a certain sample size, there seems to be diminishing returns in collecting more data.

In order to understand why, and when, this intuitive understanding of sampling works it is helpful to turn the problem on its head, and ask when, and how, might sampling fail? In other words, what are the circumstances under which the estimate we get from our sample will be a poor approximation to the true 'all population' value?

One of the most damaging sampling failures occurs when the sample we select is systematically different to the population we are trying to represent. So, if our sample for measuring blood pressure included a higher percentage of young people than the population as a whole, then we can be reasonably sure our sample will give a blood pressure average that is too low (since blood pressure is known to increase with age). So, as a primary rule, the sample must be a fair representation of the population we are interested in.

The second potential failure is if we take too small a sample. To get an estimate of average blood pressure that is close to the population average, it seems clear that we can be more confident in our estimate if we take a sample of 10,000 than if we take a sample of 100. One of the major contributions of sampling theory is that, not only can we demonstrate that our intuition here is correct, we can also quantify how much more confident we can be with the larger sample size. Hence the ubiquitous 'confidence interval'. 'Confidence' in this sense, is written in terms of statements about how far away from the true estimate we think our sample estimate might possibly be: even with larger sample sizes we can never be absolutely sure we will match the population value, but there is a smaller probability of being a long way off this figure than there is with a smaller sample size.

The third factor that can make an impact is variability in the population. To take an example, imagine we are trying to estimate average height and average weight among adult women using a sample survey. Then, it is known that the vast majority of the female adult population is of a height within the range 145 to 175 cm, the standard deviation being about 6cm and the mean 161cm. Whereas weight is about twice as variable, the range being about 45kg to 100kg, with a standard deviation of about 14kg (mean=63kg). Now, if you imagine the scenario where a sample is selected but we are very unlucky in our selection and just happen to over-sample taller women. Then it is clear that even under the very worst case scenario, where the sample *only* includes women who are at the top of the height range (no pun intended) the largest mean our sample can give us is about 175cm. In which case the difference between the true mean and the sample mean is 14cm. In other words, the largest possible margin between the sample mean and the population mean is 14cm. For weight however, if we were unlucky enough to select a sample with only the very heaviest women, then we would have a sample mean of about 100kg. So, for weight, the largest possible margin between the sample mean and the population mean is 37kg. This is a margin more than twice as large as the worst error margin for height. It follows from this type of argument, that if we want our sample to give an estimate that is close to the population value, we need to take into account how much variability there is in the variable

we are trying to measure. All else being equal, the greater the variability the larger the sample size we need[10].

So, in summary, samples will give estimates reasonably close to population values if the sample we select has no systematic bias, and if the sample size is large enough. Furthermore, the more variable the population is the larger the sample size that will be needed.

One way to think about sample design is the art of minimising the chance of getting a skewed, or extreme, sample, whilst minimising the survey cost. The main factor that the sample designer can adjust is the sample size. But there are many other factors that play a part: sampling frame, stratification, clustering. All of these are discussed in the chapter, but we begin with the most fundamental, namely the *sampling frame*.

5.2 Sampling frames

All samples involve, at least conceptually, the notion of a sampling frame, the frame being a list that covers (hopefully) all the 'elements' (that is 'persons' usually) in the population that we are sampling from.

The ideal sampling frame is a straightforward list of the elements we are trying to sample. So, for a population sample, a comprehensive list of members of the population would be ideal. And for a household survey a comprehensive list of households would be ideal.

In practice the ideal sampling frame hardly ever exists. In the UK, for instance, there is no population register that can be used for sampling purposes. The closest is the electoral register, but because inclusion on the electoral register is voluntary large sections of the population are excluded. Furthermore, it is now possible to include oneself on the electoral register but to refuse for your name to be used for any other purpose. This means that the electoral register tends to give biased samples.

The default alternative to the electoral register in the UK is the small-user Postcode Address File (PAF) which is the Post Office's list of all addresses in the UK, which receive less than 25 items of post per day. The list is primarily residential addresses (about 94%), and the list covers approximately 99% of all residential addresses.

The PAF reasonably closely approximates to a sampling frame of households, which makes it ideal for surveys of households (once non-residential and unoccupied addresses have been screened out). It can also be used as a sampling frame for individuals, but to do so the sample has to be selected in two stages: a sample of households (first stage) from which a sample of individuals is then selected (second stage). Depending upon the needs of the survey, it is usual to either select all

[10] This discussion assumes that you wish to have an estimate of mean weight that is of similar precision to your estimate of mean height. In practice this may not be the case, and your sample size calculations would need to take this into account.

adults at a household for an interview, or to select just one adult at random. How this choice is made is described in section 5.5 below.

Use of the PAF is now very well established for face-to-face interviews of the general GB population. It does have one major shortcoming in that it tends to exclude institutions (such as care homes or university halls of residence) and most GB surveys have now become surveys of the 'general household population' rather than the 'general population'. To include institutions special boost samples have to be selected.

It is worth noting that using PAF as a sampling frame for individuals is relatively unproblematic *as long as* the survey is being carried out face-to-face, because the interviewer can do the selection of an individual within the household. For postal surveys of individuals PAF sampling is really not possible because there is no good way of controlling who completes the questionnaire. Instructions, such as asking a household to select the person who has most recently had a birthday, or to distribute a questionnaire to all household members, are not adhered to strictly enough to be robust. So for population based postal surveys the default-sampling frame is still the electoral register, even though its problems are well known.

The third 'sampling frame' of the general population worthy of mention is more virtual than actual, namely the list of all possible residential telephone numbers. Although undoubtedly BT holds such a list, only the directory list is released to researchers: ex-directory numbers are excluded. Nevertheless, sampling methods do exist based on the entire list. In essence this involves selecting 11 digit numbers at random, the first 7 digits being randomly selected from the published list of prefix numbers (such as 020 8693 XXXX) and the last four digits being generated entirely at random. The selected numbers are then dialled and numbers not in use, and business numbers, are screened out to leave just residential numbers. This method of sampling is known as Random Digit Dialling (RDD).

The above discussion probably suggests that, for general population samples, the choice of sampling frame is largely determined by survey mode (PAF for face-to-face interviews, ER for postal surveys, and RDD for telephone surveys). This *is* largely the case, but the issues get more complicated when the aim is to target sub-samples of the population. For instance, a face-to-face interview survey of children could use a sample based on a PAF sample with screening out of adult-only households. But this would involve interviewers approaching more households than will be included in the survey. An alternative, more cost-effective approach, *might* be to use an alternative-sampling frame such as child benefit records. Or, depending on the nature of the survey, it might be possible to sample via schools (i.e. select a sample of schools and then a sample of children within schools). A fourth option would be to sample children from another survey: for instance the Health Survey for England might be used as a sampling frame for a follow-up survey of

children. In practice, when there are various sampling options some will be ruled out quite easily, but others will involve hard choices between cost-effectiveness and potential biases introduced by using imperfect sampling frames or frames where the sampling mechanism is likely to create non-response problems.

5.3 Sample size

The main decision needed in deciding on a sample design is sample size. To decide on this a number of questions need to be decided on:

1. What are the key estimates for the study?
2. How precise do those estimates need to be? (i.e. what size of standard error or confidence interval can be tolerated?)
3. Are there key sub-groups for which separate estimates will be needed?
4. Does the survey need to be large enough to detect change over time between surveys, or differences between key sub-groups?[11].

The basic formula that survey statisticians use to determine sample size is the 'standard error for a mean from a simple random sample survey':

$$sderr(\bar{x}) = \sqrt{\frac{S^2}{n}}$$

Where S^2 is the population variance and *n* is the sample size.

The 95% confidence interval for the mean is then calculated as

$$\bar{x} \pm 1.96 sderr$$

Note that, as per the discussion above, the standard error increases as the population variance increases, and decreases as the sample size increases.

In surveys being carried out for the first time S^2 has to be (gu) estimated. Life is simpler if we are interested in percentage rather than means because then S^2 becomes $p(100-p)$ where p is the percentages. It is usually easier to estimate a value of p in advance than it is to estimate S^2 for a mean.

[11] Note, this last requirement needs a power calculation rather than a standard sample size calculation.

In practice simple random samples are relatively rare in practice and survey statisticians use an amended version of the basic formula:

$$sderr(\bar{x}) = deft\sqrt{\frac{S^2}{n}}$$

The multiplier 'deft' in the above equation is the 'design factor'. The deft is essentially a factor that adjusts the standard error because of design features. These features include:

1. Stratification of the sample either to guarantee that sub-groups appear in the correct proportions (proportionate stratification) or to over-sample sub-groups (disproportionate stratification).
2. Weighting of the sample to adjust for non-response
3. Clustering of the sample.

Each of these is discussed below, beginning with the last on the list above: clustering.

5.4 Clustering

A 'clustered' sample is defined as a sample that is selected in two or more hierarchical stages, different 'units' being selected at each stage, and with multiple sub-units being selected within higher order units. A few examples will help to clarify this:

- A sample of children is selected by (a) sampling schools and then (b) selecting children within schools. This is a two-stage clustered sample, the clustering being of children within schools.
- On a general population survey a PAF sample is used to generate a sample of households. Within each household up to two adults are selected at random. This is a two-stage clustered sample, the clustering being of adults within households. Note that, had the instruction been to select just one adult per household, this would *not* be described as a clustered sample, because there would no clustering of the adult sample within a smaller number of households.
- The most common design for PAF samples is, at the first stage, to select a random sample of postcode sectors[12]. Then, at the second stage, households are selected within these postcode sectors. And then, at a third stage, individuals might be selected within households. Under this design adults are clustered within households (assuming more than one adult is

[12] As an example, the postcode EC1V 0AX is part of the postcode sector EC1V 0.

selected per household) and households are clustered within postcode sectors.

Clustering, or multi-stage sampling, is adopted on surveys for a number of reasons. The two main reasons are:

- because the sampling frame units cover two or more survey units and so clustering is the only practical way of selecting a sample of the units required
- to divide the sample into manageable workloads for interviewers.

5.4.1 Clustering because of the sampling frame

If the sampling frame to be used for a survey consists of units larger than the survey units, then it is very common to use a clustered sample design. Generally speaking the more survey units each sampling frame unit covers the more clear-cut the case for clustering will be.

For instance, if a survey of employees is to be carried out using a sampling frame of business establishments, then the most cost-efficient solution is almost bound to be to select a sample of establishments and then to select a sample of employees per establishment. The 1998 Workplace Employee Relations Survey adopted this design, with a sample of about 2000 establishments being selected at the first stage, and then 25 employees being selected per establishment at a second stage, giving a total sample size of about 50,000. Although it would have been possible to take an unclustered sample by selecting 50,000 establishments and then one employee per establishment, the fact that this would involve negotiating access via such a large sample of establishments renders this approach completely impractical.

An instance where the merits of clustering are less clear-cut is the selection of individuals within households. On some general population samples several adults are selected per household (to give a clustered sample) whereas in other surveys just one adult is selected per household (giving an unclustered sample).

5.4.2 Clustering to give manageable interviewer workloads

Almost all face-to-face interview surveys of the GB population use geographical clustering at the first stage of sample selection. The usual procedure is, as was noted above, to select a sample of postcode sectors and then to select PAF addresses within these postcode sectors. For instance, the main Health Survey for England sample is selected by firstly selecting 720 postcode sectors and then 20 addresses per sector.

The rationale behind this approach is that the sample of addresses in each postcode sector becomes the workload for one interviewer. So an

interviewer is given 20 addresses to interview at, and these are located in a relatively small geographical area. The interviewer does not then have large distances to travel between sampled addresses, and the costs of the survey are reduced.

The observant reader may notice that, on the face of it, this sampling procedure (whereby equal sample sizes of addresses are selected per sector) means that people who live in large postcode sectors (i.e. postcode sectors with a large number of addresses) will be under-represented in the sample. This is avoided by over-sampling large sectors at the first stage of selection. More precisely, sectors are selected with probability proportional to their address count (probability proportional to size sampling). Taking an equal sample per sector at the second stage then counterbalances this over-sampling, to give an equal probability of selection per address overall.

5.4.3 The impact of clustering on standard errors

The main objection to clustered samples is that they tend to give estimates with larger standard errors than unclustered samples. That is they give a deft greater than one. The reasoning here is that the more the sample is clustered the greater the chance we have of drawing a sample that is extreme. For instance, imagine a scenario where we are selecting a sample of 1000 people. Then if we choose to select the 1000 people by selecting just 10 postcode sectors and 100 people per sector, then if we are unlucky enough to select one or two sectors than are outliers in some sense then we will get a sample mean that is quite different to the population mean. If, instead, we choose to select 100 postcode sectors and 10 people per sector, then the impact on sample estimates of, by chance, selecting one or two outlier postcode sectors will be much smaller. Under this less clustered design we can be sure the sample will give estimates that will be reasonably close to the population mean. However, we could be even more confident if we unclustered the sample even further, by taking, say, 500 postcode sectors and just two people per sector.

From this example it is hopefully clear that the more the total sample is spread across clusters the lower the chance of taking an extreme sample and the lower the standard error. This translates into: for a fixed sample size, the smaller the sample size per cluster the smaller the standard error. Note however, that in the example above clustering will only be a problem if there is a risk of selecting a non-representative sample of postcode sectors (described above, as over-representing outliers). This can only happen if postcode sectors differ from each other, that is, if there is between-sector variance. If all sectors are the same then no matter what sectors are selected the survey estimates will be the same. So, standard errors associated with clustering increase *if* there is between-cluster variance, and as the sample size per cluster increases.

On the other hand, as was noted earlier, clustering of a general population sample within postcode sectors tends to reduce interviewing costs because it reduces travel costs for interviewers. So, decisions on the extent of clustering involves judgement about how standard errors can be minimised for a fixed survey budget – a clustered sample may give larger standard errors than a simple random sample, but a larger sample size will be affordable with a clustered sample, so the impact of clustering can usually be more than offset. For most face-to-face interview surveys a clustered sample will be the most cost-effective.

To model the effect of clustering survey statisticians make use of a crude estimate of the design factor

$$deft = \sqrt{(1 + (m-1)roh)}$$

Where m is the average sample size per cluster and roh is a measure of the relative between-cluster variance[13]. Roh values differ from estimate to estimate, but tend to be highest for variables that are very geographically clustered (such as tenure, and to a lesser extent deprivation). Roh is very small for variables that are roughly constant across clusters (e.g. sex or age). One of the difficulties of survey design is that it is necessary to estimate roh in advance of doing a survey – unless the survey is a repeat of an earlier survey this is more of an art than a science.

It is worth noting that the value of roh will tend to vary depending upon the geographical definition of the clusters. Generally, the smaller the cluster, in geographical terms, the larger roh will be. This is the reason survey organisations tend to use postcode sectors, which cover areas of about 2300 households, rather than smaller geographical areas such as enumeration districts.

5.4.4 Clustering within households

Using PAF as a sampling frame for general population surveys means that a decision is always needed on whether to select one adult per household or more than one adult per household. For instance, the Health Survey for England selects all adults per household, and the annual British Social Attitudes survey selects just one.

The arguments for and against are:

1. Selecting more than one adult per household introduces a second tier of clustering into the sample. This will tend to increase standard errors.

2. But selecting just one adult per household means that adults from larger households are under-represented in the final

[13] Roh varies from 0 to 1, being zero when there is no between-cluster variance, and 1 when all the variance between population members is between clusters, and there is no variance within clusters.

sample. This has to be adjusted for by weighting the final sample (see section 5.8 below). This will also increase standard errors.

So both options increase standard errors. But

3. Selecting more than one adult per household means that fewer households need to be included in the final sample. This reduces survey costs.

4. However selecting more than one adult per household puts a lot of burden on the household and this can increase non-response.

In practice it is usual to select more than one adult per household *if*

a) this is not expected to excessively burden the household; and

b) household members are not expected to be too homogeneous in terms of the things the survey is trying to measure (which would give large roh values); and/or

c) most of the analysis will be done within male/female sub-groups (which has the effect of giving relatively unclustered subgroups).

5.5 Stratification

Alongside decisions on how to cluster a sample, decisions also need to be taken on stratification.

Stratification essentially means dividing the sampling frame into groups (strata) before sampling. A simple example would be to take a sampling frame of, say, business establishments and then to sort them into size strata before sampling. The sample would then be described as a sample stratified by size. If a list of the general population was available that had age and sex recorded, then it would be possible to divide the list into age and sex strata before sampling to give a sample stratified by age and sex.

There are two methods of stratified sampling: proportionate and disproportionate. In a proportionate stratified sample the sampling frame is divided into strata but the same sampling fraction is applied per stratum. This means that each stratum is sampled from in its correct proportion. In a disproportionate stratified sample the sampling fraction differs between strata. This means that individuals from the strata with the highest sampling fractions will be over-represented in the sample. Disproportionate sampling is generally used when there is a need to boost the sample size within a particular stratum or strata.

Proportionate stratified sampling is a more powerful tool than it may appear to be at first glance. The main advantage it has over simple

unstratified random sampling is that it *guarantees* that the sample drawn matches the sampling frame in terms of the strata. So, for instance, if the sampling frame is stratified by age and sex and a proportionate sample selected, then the sample will match the sampling frame in terms of the age-sex distribution. In other words the age-sex distribution is controlled. If the survey data happens to be correlated with age and/or sex, then it follows that by stratifying the sample by age and sex there is less risk of drawing an extreme sample which gives survey estimates far removed from the population values. In other words, stratification reduces standard errors.

The degree to which standard errors are reduced depends on how closely associated the strata variables are to the survey estimates. For example, in a health survey measuring physical health, stratification by age would be very powerful because physical health is so closely related to age. Stratification by sex would be useful, but less so, because there is a weaker relationship between sex and physical health than there is between age and physical health. In contrast, for a survey of mental health, sex would probably be the better stratifier, because there is a stronger relationship between sex and mental health then there is between age and mental health.

Even though stratification is very useful in minimising standard errors, its use is often restricted by the fact that it is only possible to use variables as stratifiers if they appear on the sampling frame. So, in practice, age and sex stratification for general population samples is very rare because none of the usual sampling frames include age and sex. For PAF based samples the possible stratifiers are all geographical indicators, such as location, and characteristics of postcode sectors as derived from the most recent census (such as percentage of households with an unemployed head of household). There are no good stratifiers at the level of individual addresses.

Typically, what happens on PAF samples is that PAF is divided into regional strata (typically Government Office Region) and then, within regions, postcode sectors are divided into strata using one or two variables thought to be reasonably closely related to the survey subject matter. So a survey of income might use an area-level deprivation index as a stratifier for postcode sectors, and a travel survey might use an urban/rural stratifier.

We noted earlier that clustering of PAF samples within postcode sectors tends to increase standard errors, giving a deft of greater than one. The effect of selecting the sectors within strata is to reduce the impact of the clustering (in the sense that it reduces the risk of selecting a skewed sample of sectors) with the result of reducing the deft. The experience on most surveys is that, with careful selection of stratifiers the deft can be quite significantly reduced, but it tends still to be greater than one. In other words, a clustered sample with clusters selected using stratification will still give larger standard errors than an unclustered

random sample of the same size. Nevertheless, stratified clustered samples have been found to be the most cost-effective sample design for face-to-face interview surveys: for the same cost, larger sample sizes can be achieved than with unclustered designs, and the increased sample size more than offsets the design factor due to clustering.

5.6 Quota sampling

Most government-sponsored surveys use *random* sampling methods (or probability sampling as it is often called). On a PAF-based clustered sample this means that the postcode sectors will be selected within each *stratum at random* (albeit with probability proportional to size), a random sample of addresses will be selected within selected sectors, and assuming there is some selection within households, individuals will be selected from within a household using a random sampling method. The aim of random sampling is to avoid any self-selection bias in the sample, whereby areas are selected because interviewers prefer to work there, and individuals are selected who are more willing or able to take part in surveys. In strict random samples, once the sample is selected there can be no deviation from the sample. So those who refuse to take part become 'non-respondents' and, importantly, no attempt is made to replace non-responders with others willing to respond.

The main alternative sampling method is *quota* sampling, a method used fairly extensively in market research. Quota samplers allow substitution of non-respondents with others willing to respond, but they make considerable use of stratification principles to ensure the final sample reflects the population at least on some key variables. For instance, a quota sampler might use population estimates of the numbers within each combination of age group, sex, and social class to decide what numbers are needed within each of these combinations for a survey. Interviewers are then given quotas based on these numbers that they are asked to achieve.

The theory behind quota sampling is that, as long as the variables used to determine the quotas are selected carefully enough, then the fact that within a quota cell there is no attempt at random sampling, is not important. The survey estimates will still be unbiased. The underlying assumption is that, within a quota cell, those who take part in the survey have the same characteristics, attitudes, behaviours etc. as those who do not take part. The skill is to find variables for the quota that control for most of the survey variability between individuals – if this can be done then the assumption that responders are similar to non-responders becomes more credible.

Of course, even in random sample surveys a fairly high percentage of those selected will refuse to take part, so, in producing survey estimates assumptions have to be made that responders have similar characteristics to non-responders. So, in that respect, random and quota sampling are similar. One way to think about the difference is that

random sample surveys tend to achieve response rates in the range 60-80%. Quota sample surveys can be thought of a random sample surveys but with much lower response rates, perhaps in the order of 20-30%. Quota samplers have, in fairness, managed to demonstrate empirically that quota samples can give very similar estimates to random samples, even on surveys where non-response bias would be expected to be a major problem. Nevertheless, random sampling is still the lower risk approach, and the UK government has tended to remain faithful to random sampling even though quota sampling methods are considerably cheaper and faster.

5.7 Sampling special populations using screening

It is often the case that surveys have a particular focus on sub-groups of the general population rather than on the whole population. Some examples include:

- the 2003 Health Survey for England which had a focus on child health and included a boost sample of children
- the 1996 British Crime Survey which included a boost sample of people from minority ethnic groups
- the Low Income Diet and Nutrition Survey which specifically includes only low-income households.

The usual approach to selecting the sample in these instances is to select a large PAF based sample, to carry out a short screening survey at all households (which might be done on the doorstep if it is very short) and to carry out a full interview only in households with the relevant people.

The main issue here tends to be cost, since, in general, interviewers will have to screen out more households than they include. Boost samples of children are relatively inexpensive because a fairly large percentage of households have one or more children (about 30%). Boost samples of minority ethnic groups are far more expensive because interviewers have to screen at a large number of addresses to achieve the final sample. In fact, minority ethnic boost samples tend to include over-sampling (i.e. disproportionate stratification) of postcode sectors where the percentage of ethnic groups is higher than average in an attempt to improve interviewer screening rates and, hence, to reduce survey costs.

The success of screening largely depends on our ability to ask screening questions quickly, and at the start of interviews. This is pretty straightforward when screening for children and minority ethnic groups – this can be done on the doorstep as long as the interviewer is careful to explain why the questions are being asked. Other screening questions are harder to ask up-front, and surveys have on occasion used proxy measures. An example of this is the Low Income Diet and Nutrition Survey where it was not considered feasible to ask detailed questions about income on the doorstep, so, instead, a proxy indicator of relative

deprivation was used. This included questions on tenure, car ownership, employment, lone parent status, and benefit receipt.

5.8 Survey Weighting

In most surveys it will be the case that some groups are over-represented in the raw data and others under-represented. This might be because of the sample design, primarily because of the use of disproportionate stratification or boost sampling, or because of sampling features that lead to unequal probabilities of selection, such as selecting one person per household on PAF samples. Alternatively some groups may be over- or under-represented because of non-response patterns.

These mis-representations are usually dealt with by weighting the data. The idea behind weighting is that members of sub-groups that are thought to be over- or under-represented in the survey data are each given a weight. Over-represented groups are given a weight of less than one; under-represented groups are given a weight of greater than one, the weight being calculated in such a way that the weighted frequency of groups matches the population. All survey estimates are calculated using these weights, so that averages become weighted averages, and percentages become weighted percentages, and so on.

The calculation of the weights for a survey is rarely a straightforward business. Weights for disproportionate sampling are relatively non-controversial, but weights to adjust for non-response biases are largely dependent upon judgement, and it is likely that no two analysts would ever calculate exactly the same set of non-response weights. Nevertheless, following the GSS task force on weighting some standardisation is now coming into play. The main principles are:

- non-equal probabilities of selection (including disproportionate stratification) is dealt with by applying weights proportional to the inverse of the probability of selection;
- at a minimum non-response is dealt with by weighting survey data to published distributions by age, sex and region.

The actual means of calculating these non-response weights can differ from survey to survey, but the most commonly used method now is 'calibration weighting' (see Box 5.1 for an explanation of calibration weighting and how the method has been used in the British Crime Survey, a household survey of crime victimisation).

Box 5.1 Calibration weighting and the British Crime Survey (Source: Crime in England and Wales 2002/2003, Home Office Statistical Bulletin 07/03)

The Office for National Statistics (ONS) recommended that the calibration weighting method be adopted in the British Crime Survey (BCS). The weighting is designed to make adjustment for known differentials in response rates between different age by gender subgroups and households with different age and gender composition. For example a 24 year-old male living alone may be less likely to respond to the survey than one living with a young partner and a child. The procedure therefore gives different weights to different household types based on their age/sex composition in such a way that the weighted distribution of individuals in the responding households matches the known distribution in the population as a whole (based on population estimates provided by ONS). The weights are generated using an algorithm (CALMAR) that minimises the differences between the weights implied by sampling and the final weights subject to the weighted data meeting the population controls.

The calibration weighting method is now used on the General Household Survey (ONS), the Expenditure and Food Survey (ONS and DEFRA), the Family Resources Survey (DWP), Family and Children's Survey (DWP) the Labour Force Survey (ONS), and other surveys.

By and large weighting of survey data tends to increase the standard errors of estimates. A key issue for non-response weighting is whether the reduction in survey bias is adequate compensation for the increase in standard errors. In some surveys it will be, in others not.

5.9 Further reading

Barnett, V. (2002) *Sample Survey Principles & Methods*. Arnold (3rd Edition).

Barton, J. (1996) Selecting Stratifiers for the Family Expenditure Survey (FES), *Survey Methodology Bulletin*, 39: 21-26.

Cochran, W. G. (1977) *Sampling Techniques*. Wiley (3rd edition).

Elliot, D. (1991) *Weighting for Non-Response: A survey researcher's handbook*. OPCS.

Greenfield, T. (ed.) (1996) *Research Methods: Guidance for Postgraduates*. Arnold.

Kalton, G. (1983) *Introduction to Survey Sampling.* Sage.

Kalton, G. (1983) *Compensating for Missing Survey Data.* Institute for Social Research, University of Michigan.

Kish, L. (1992) Weighting for Unequal P., *Journal of Official Statistics*, 8: 183-200.

Kish, L. (1965) *Survey Sampling*. Wiley.

Lynn, P. & Lievesley, D. (1991) *Drawing General Population Samples in Great Britain*. SCPR.

Lynn, P. & Taylor, B. (1995) On the bias and variance of samples of individuals: a comparison of the electoral registers and postcode address file as sampling frames, *The Statistician*, 44:173-194.

Moser, C. A. & Kalton, G. (1971) *Survey Methods in Social Investigation*. Gower (2nd edition).

Stuart, A. (1984) *The Ideas of Sampling*. Griffin.

The Magenta Book: guidance notes for policy evaluation and analysis

Background paper 6: how are the data collected?
Data collection & survey design

Published March 2004
Updated October 2007

Government Social Research Unit

HM Treasury
1 Horse Guards Road
London SW1A 2HQ

6 How are the data collected?

6.1 Methods of data collection

When deciding on how the data should be collected (the methodology) it is useful to consider the types of questions the research is attempting to address. The following illustrates the range of questions one might wish to ask about a particular issue, such as teenage pregnancy or homelessness.

1. How many people are in this situation, are affected by this problem, have been helped by this initiative?
2. How prevalent is this problem?
3. Which groups are most affected by these issues/ are most at risk?
4. How much of a difference does the initiative/programme make to the prevalence of these problems?
5. Why are people affected by this problem?
6. How do people end up in this situation?
7. How does the initiative / programme work?
8. Why does it work, not work?
9. What do people think about the intervention?/How could it be improved?

If we are principally concerned with knowing the answers to questions 1 to 4 then a quantitative methodology would be more appropriate. If, however, we are principally concerned with knowing the answers to questions 5 to 9 then a qualitative methodology may be more suitable.

6.1.1 Strengths and limitations of qualitative and quantitative methods

The strengths and limitations of qualitative and quantitative[14] methods are outlined in Table 6.1.

[14] Quantitative here refers to sample-survey methods, although there are many other study designs that use quantitative methods (e.g. randomised controlled trials).

Table 6.1 Summary of advantages and limitations of qualitative and quantitative (survey) methods

	Qualitative	Quantitative
Strengths	Flexible	Produces statistical data
	Enables exploration of the meaning of concepts, events	Where random probability samples are used, survey estimates can be defined within specified bounds of precision
	Produces valid data as issues explored in sufficient depth to provide clear understanding	Can measure the extent, prevalence, size and strength of observed characteristics, differences, relationships and associations
	Enables study of motivations and patterns of association between factors	Can determine the importance of factors in influencing outcomes
	Provides a detailed understanding of how individuals interact with their environment, cope with change etc.	Uses standardised procedures and questioning, enabling reproducibility of results
Limitations	Sample sizes are often small	Can be costly, particularly if population rare or 'hard to reach'
	Need to be able to anticipate factors associated with issues to be studied, to design 'good' sampling strategy	Sampling frame may not be available
	Interviewing methods rely on respondents being reasonably articulate	Structured interview hinders detailed exploration of reasons underpinning decisions or views
	Analysis of data to generate findings is not always transparent or replicable	Standardised questionnaire design and administration means there is little flexibility to be able to deal with respondents' misunderstanding the question (or its intention), leading to problems of validity
	Generalisability of findings can be an issue	Requires key concepts to be clearly defined and translated into meaningful survey questions. 'Fuzzy' concepts are difficult to measure

6.2 Combining qualitative and quantitative methods

Qualitative and quantiative methods can be combined and this is a useful strategy for both measuring the topic of interest and providing a detailed understanding of its nature or origins.

Table 6.2 illustrates how the different types of evidence obtained from quantitative and qualitative methods would contribute to a study about bullying among school children and the effectiveness of an intervention.

Table 6.2 Contribution of qualitative and quantitative evidence to answering research questions: bullying among school children#

Qualitative methods investigate/ understand	Quantitative methods measure
The nature of different forms of bullying	The extent to which different forms of bullying exist
The experience of being bullied and being a bully	The characteristics of those bullied and of bullies
The events leading to bullying/ the circumstances in which it occurs	Factors associated, statistically, with being bullied/ being a bully
Why bullying continues	Characteristics/ circumstances that correlate with length of time being bullied/ bullying
Appraisal of any interventions experienced	Extent to which schools have anti-bullying policies
Influential factors in bringing periods of being bullied/ being a bully to an end	Extent to which policies have an impact on levels of bullying in school
Suggestions / strategies for supporting those bullied/ bullies	Prediction of future levels of bullying
	Prediction of resources required to deal with bullying effectively

Based on Ritchie, 2003.

When deciding whether to combine qualitative and quantitative (survey) methods of data collection, it is important to consider what types of evidence or information are required and at what stage in the research process this evidence or information will be needed.

To get the most out of combining qualitative and quantitative methods requires:

- a clear set of research questions;
- a reasonable timeframe for the research;
- close working relationships between the qualitative and survey researchers (if they are different people); and
- sufficient funding to allow this close working.

6.3 Quantitative (survey) methods of data collection

There is a range of different types of data collection methods that can be employed in collecting quantitative survey data. These are outlined in Table 6.3.

Table 6.3 Types of data collection methods

Interviewer-administered methods	Self-completion methods
Face-to-face	Postal
Telephone	Web/email

In broad terms there are three sets of factors that will influence the decision over which data collection method to employ:

- survey administration and resource issues;
- questionnaire issues; and
- Data quality issues (Czaja and Blair, 1996).

The differences between face-to-face, telephone and postal surveys, in terms of these factors, are outlined in Table 6.4.

Table 6.4 Summary of the strengths and weakness of different modes of data collection#

Design parameter	Face-to-face	Telephone	Postal
Cost of data collection	Usually most expensive method	Usually around 50-70% of face-to-face cost for same interview	Relatively cheap (but q'naires need to be kept short and simple)
Amount and type of resources required	Specialised fieldworker skills and field-force management resources needed	Specialised interviewer skills and management resources needed	For samples < 1,000 normal office resources suffice
Timetable considerations	May require several months unless respondents are easily accessible or 'captive'.	Usually the fastest mode of data collection, but depends on respondent availability	With response reminders, may require several months
Operational control	Best for control of field sampling and data collection	Good for interviewer supervision, but respondent tolerance may be limited	Few means of controlling how q'naires are completed
Amount/complexity of data to be collected	Best/mandatory for long and complicated questionnaires	Limitations on length and data collection complexity compared with face-to-face	Weaker for groups with poor literacy or motivation, but can be good for experts
Likely quality of the data	Best for complex topics and issues. Computer assistance improves quality. May incur interviewer effects	Good for simple factual and attitudinal questions. Computer assistance improves quality. Interviewer effects less likely	Worst for missing data, routing errors, misunderstandings
Statistical efficiency	To reduce fieldwork costs less efficient clustered samples needed for national surveys	Does not require clustered samples, but may have sampling problems	Does not require clustered samples
Expected response rate	Usually gets highest rate	Likely to be 10-40% lower than face to face	Usually lowest rate. Can be well below 50% for less literate/motivated

Based on Lynn & Thomas, 2003.

6.3.1 Combining methods

There are times when it may be appropriate to combine different methods of data collection, for example to:

- save money;
- improve geographical coverage;
- overcome sample frame bias;
- improve data quality, such as response rates or item non-response;
- speed up data collection; or
- overcome resource problems, such as a lack of face-to-face interviewers.

Despite these advantages there are potential pitfalls to combining data collection methods.

- Development time may need to be extended, as two or more data collection instruments will need to be designed.
- Survey management costs will be increased, as different groups of people will receive different treatments (although these may be offset against savings in the overall cost of the survey).
- Keeping track of the outcome (interview, refusal, non-contact etc) for each case will be required at each stage of the data collection process, particularly if the design involves a follow up of non-responders using an alternative data collection method.
- Particular care will be required to avoid data being lost or duplicated, as a result of having to stitch together data collected from different sources.
- Mode effects can impact on the reliability and validity of the data collected.

6.4 Computer Assisted Survey Information Collection (CASIC)

It is now commonplace for large-scale face-to-face interviews to be conducted using Computer Assisted Personal Interviewing (CAPI), whereby the questionnaire is a computer program loaded on to a lap-top computer that an interviewer takes out into the field. Respondents' answers are entered into the laptop and interviews transmitted back to the office via modem.

Paper and Pencil Interviewing (PAPI) methods, in contrast, require questionnaires to be posted back to the office, where the information has to be converted into an electronic format, either by being keyed or scanned, which takes longer.

Computer Assisted Telephone Interviewing (CATI) is widely used. Here the questionnaire is accessed via a computer terminal located in a centralised telephone unit.

Computer Assisted Self-Interviewing (CASI) enables respondents to complete a self-completion questionnaire using a laptop computer. Audio-CASI enables respondents to hear the questions rather than relying on them being read.

CASIC methods offer many advantages for surveys. They:

- Automatically direct the interviewer or respondent to the appropriate questions based on answers given earlier in the questionnaire. It therefore prevents interviewers (or respondents, if CASI) from making routing errors.
- Allow complex routing that would be impractical on paper questionnaires and potentially error-prone (e.g. missing data, answers to inapplicable questions).
- Interviewer can concentrate on the actual questions and respondents' answers, if CAPI or CATI, or if CASI the respondent can concentrate on answering the questions.
- In CAPI and CATI warnings can be triggered if improbably extreme values are entered, or if there is an inconsistency between answers at different questions. Substantial timesavings after the completion of the fieldwork (although more time is needed in the beginning to set up CASIC questionnaires compared to paper ones).
- Data are entered directly into a computer during the interview, so data entry as a separate task is eliminated.

6.4.1 Web and email-based data collection

Web and email-based methods of data collection are not widely used in social research at present. This is mainly because of difficulties over sampling; specifically about being able to select random probability samples for general population surveys.

The advantages of CASIC methods can be realised, such as a reduction in routing errors and speed of data transfer. However, to achieve this a significant programming effort is required, using a Web-based language such as Java, which in turn is costly and time-consuming.

6.5 Survey instruments

There are a number of different types of survey instrument that can be used to collect the information required to answer the research questions. These include:

- structured questionnaires;
- diaries;
- measurements (e.g. height, weight);
- tests (e.g. reading, memory); and
- observations (e.g. quality of house conditions).

Questionnaires collect information by means of pre-scripted questions. The questionnaire can be either administered by an interviewer or completed by the respondent. In the case of the former, the question order is predetermined. Questionnaires can collect factual, behavioural and attitudinal information as well as measuring respondents' knowledge, although the latter can only be reliably collected if an interviewer administers the questionnaire or the respondent completes it in a controlled environment. The mode of data collection can influence the reliability and accuracy of the information obtained. For example, the accuracy of information on 'sensitive' behaviours, such as drug taking, may be influenced by whether the data are collected by an interviewer or using a self-completion method.

Diaries can be used to collect detailed behavioural information, for example on diet, travel or time-use. Diaries allow information to be collected prospectively, that is at the time of the event. They are a form of self-completion questionnaire, with respondents being asked to record details of the behaviour of interest over a specified time period. In this way it is hoped that details of respondents' usual behaviour are captured. Diaries can capture much more detailed information about behaviour than is often possible in a questionnaire, and can be used alongside structured questionnaires.

Measurements can be taken to collect factual information such as respondents' height, weight, blood pressure, blood iron levels and so on. As with diaries, these measurements can be collected in conjunction with information obtained from a questionnaire (and diary). Protocols need to be developed to ensure these are taken in a standardised way. Ethical approval may be required.

Tests can be administered, as part of the survey interview process, to measure respondents' ability to perform certain tasks, such as reading or walking. Such tests are often standard assessment tools that have been developed for a particular setting, such as a clinical

or educational assessment in a hospital or school. As with the collection of measurements, protocols will need to be developed that ensure the tests are administered in a consistent way and that they can be administered (reliably) in a survey interview situation.

Observations can be made of factual information, such as the condition of the respondent's accommodation. Observers need to be carefully trained to record information in a consistent way. Observational data can be collected alongside other types of information to provide a more detailed picture of respondents' circumstances.

6.6 Survey questions

Survey questions can be asked that seek different types of information:

- factual;
- behavioural;
- attitudinal; and
- knowledge.

Factual questions. Whilst some of these types of information can be obtained through observation or by reference to official records, they are often collected by asking people questions to obtain it. This is because surveys often offer the only practical and affordable way of collecting such information, and in some cases there is no other source or other way of measuring the attribute of interest.

Behavioural questions, as the name implies, are concerned with measuring respondents behaviour, and can be seen as being a particular type of factual question. We often want to know information about behaviour because we want to understand what people do and or what impact government policy has on them. The following are typical behaviour-type questions:

- What do people do?
- How much do they do it?
- How often do they do it?
- When do they do it?
- Where do they do it?
- Who do they do it with?
- Why do they do it?

Attitudinal questions seek to measure respondents' opinions, beliefs, values and feelings. These are subjective attributes, which cannot be

verified by reference to observation or external data sources, and as such they can be difficult to measure and validate.

The following indicates the stages involved in developing an attitude scale.

1. Determine clearly what it is you want to measure
2. Generate an item pool
3. Determine the format of measurement
4. Construct the scale
5. Evaluate the scale

Knowledge questions are used to assess what respondents know about a particular topic. An example of their use is in welfare benefits research, where they can be used to assess people's awareness of particular benefits and tax credits and the qualifying rules for them. Answers to knowledge questions can be affected by the wording of other survey questions. For example, if we want to measure what respondents' know about the qualifying rules for Working Families Tax Credit, we should ask these questions before questions that give an indication of what the rules are.

6.6.1 Open and closed questions

Closed questions constrain answers to a set of pre-scripted answer alternatives. *Open* questions have no such restrictions.

The pros and cons to asking open and closed questions are detailed in Table 6.5.

Table 6.5 Pros and cons of open and closed questions

Question type	Pros	Cons
Open	Respondents can answer in their own words	Responses have to be coded to allow statistical analysis, which is costly, time-consuming and subject to error
	Are not leading so can, potentially allow measurement of salience – how important issue is to respondent; indicate respondent's level of knowledge; indicate strength of feeling	If an interviewer-administered survey, can be difficult to get interviewers to probe consistently
	Can avoid format effects, such as primacy (the tendency to endorse the first option seen) or recency (the tendency to endorse the last option heard)	Respondents may not provide sufficient detail when answering to capture key differences
	Are required for the development of response options for closed questions	Can generate irrelevant responses
Closed	Can help with respondent-recall	If not a complete list of answer-options, can introduce bias
	Level of detail, areas of interest can be conveyed to respondent	Cannot capture detailed information, for example occupation details
	Little or no coding of answers required, thus quicker and cheaper	Categories may not be recognisable to respondents
	Less likelihood of coder bias or inconsistent probing	Answer options can influence way in which respondents' interpret question

6.7 Designing survey instruments

There are three golden rules that are useful to consider when writing survey questions:

- Can the respondent understand the question?
- Is the respondent able to answer the question?
- Is the respondent willing to answer the question?

The key rule to remember in designing questions that can be understood is to keep them simple. That way the complexities and ambiguities described above can be avoided.

6.7.1 Respondents' ability to answer the question

It should not be assumed a *priori* that respondents will have the information necessary to be able to answer the survey questions being posed. Rather we need to consider whether respondents have been exposed to the event(s) or experiences we are asking them about, and if they have, whether they will be able to remember them or not.

Even if the respondent experienced the event of interest, she or he may not be able to answer the question because:

- the information required never got stored in her or his (long-term) memory (Willis et al, 1999) ;
- the retrieval (survey) context is different to the original encoding context, so the respondent may not recognise that the event took place or be able to recall the event correctly (Tulving and Thompson, 1973);
- the item may be difficult to distinguish from other, similar, events (Anderson, 1983); or
- the memory and or the cues associated it with it, have faded over time (Sudman and Bradburn, 1974).

Table 6.6 summarises the key factors affecting recall:

Table 6.6 Summary of factors affecting recall

	Variable	Finding	Implication for survey design
Characteristic of event	Time of event	Events that happened long ago harder to recall	Shorten reference period
	Proximity to temporal boundaries	Events near significant temporal boundaries easier to recall	Use personal landmarks, life events calendars to promote recall
	Distinctiveness	Distinctive events easier to recall	Tailor length of the reference period to properties of target event; use multiple cues to single out individual events
	Importance, emotional impact	Important, emotionally involving events easier to recall	Tailor length of the reference period to properties of target event
Question characteristics	Recall order	Backwards search may promote fuller recall	Not clear whether backward recall better in surveys
	Number and type of cues	Multiple cues typically better than single cues; cues about the type of event (what) better than cues about participants or location (who or where), which are better than cues about time (when)	Provide multiple cues; use decomposition
	Time on task	Taking more time improves recall	Use longer introductions to questions; slow pace of the interview

(Tourangeau et al, 2000: 98)

6.7.2 Respondents' willingness to answer the question

Even if respondents understand the question and are in possession of the necessary information to be able to answer it they still might not provide an answer because they are unwilling to do so. Respondents need to be motivated to engage in the necessary cognitive (thinking) effort to answer the question.

Studies have identified a number of factors that can impact on respondents' willingness to answer survey questions. These include whether they perceive the request for information to be:

- legitimate;
- reasonable;
- beneficial; and

- not to have any negative consequences (Kahn and Cannell, 1957; Cannell et al, 1979).

6.7.3 Salience and length

Research has shown that the more salient respondents find the questions the more likely they are to answer them.

An excessively lengthy questionnaire can impact on survey data quality in a number of ways, including:

- lowering survey response rates;
- increasing item non-response for questions later in the questionnaire;
- increasing respondent errors.

6.7.4 Question wording and order effects

In designing questionnaires particular attention is required to the way in which questions are worded and to the order in which they are asked, as these things can have an impact on respondents' answers.

6.7.5 Question wording effects

Changes in question wording, even what appear to be small ones, can have an impact on the way in which respondents answer them.

It is often difficult to predict whether changes in question working will have any effect on response. They can be detected by conducting a split-panel experiment.

In a spilt-panel experiment one version of the question to half the respondents (group a) in the sample and the other version to the other half (group b). The allocation of respondents to group a) or b) would be random. Apart from the wording of the question all other aspects of administration of the survey question would be the same. This approach allows us to compare the results from the two question variants, and if we observe a difference in the distribution of answers between them, be certain that this difference is the result of the change in question wording alone.

6.7.6 Question order

Question order can affect the way in which survey respondents interpret survey questions and thus answer them. This is because the wording of preceding questions can help to shape the context in which respondents interpret the current question. There are two main types of context effects:

- assimilation; and

- contrast.

Assimilation effects (sometimes known as consistency effects) occur when respondents infer that the current (target) question is directly related to the proceeding (context) questions. For example, in a study conducted by Schuman and Presser (1981) the order of the following two questions was varied.

a) Do you think the United States should let Communist reporters from other countries come in here and send back to their papers the news as they see it?

b) Do you think a Communist country like Russia should let American newspaper reporters come in and send back to their papers the news as they see it?

Support for statement a) varied by 20 per cent according to whether it was asked before or after question b). The explanation for this variation, put forward Schuman and his colleagues is that when question a) is asked first, many answers reflect attitudes towards communism or Russia. When a) is asked after b), answers are based on the notion of even-handedness: American reporters should be allowed in to Russia and be able to report on the news there as they see and thus the same principal should be applied to Russian reports.

Contrast effects occur when respondents infer that the target question should be compared with the context question(s). For example, answers to general questions can be influenced by whether they are asked before or after questions about specifics.

6.7.7 Things to avoid

A summary on the "do's and don'ts" of questionnaire design is provided below:

1. Asking people for information they do not have
2. Do not ask more than one question at a time: avoid double-barrels (i.e. including two different concepts in one question)
3. Avoid double-or implicit negatives

This is particularly an issue for attitude statements involving the use of agree/ disagree response options. Consider the following example:

I am much less confident now than I used to be?

If the respondent feels more confident now than she used to then she has to disagree with the statement, which can cause problems because she has to engage in a double-negative, that is she does not agree that she feels less confident now.

4. Long lists of response choices

These can suffer from primacy or recency effects. If long lists are presented to respondents in a visual format there is a tendency for people to only look at the first few items in the list rather than read all items. If the list is read out, then respondents are more likely to only remember the last few items.

5. Beware of questions that include hidden contingencies, as these may only be relevant to a sub-section of the population. For example the question *"How often do you drive your car to work?"* clearly only applies to people who a) can drive, b) have access to a car that they can drive, c) who work and d) who do not work at home. Thus the question should not be asked of all respondents, but only of those to whom it is relevant. Filter questions should be asked, in advance of the question to establish who should be asked it. Asking questions that are not relevant to respondents is likely to annoy them, making them less likely to want to continue to answer questions and makes the answers to such questions difficult to interpret.

6. Questions that start with response choices such as "*Would you say that you often, sometimes, rarely or never [buy a newspaper]?"* Response options should always come after the question.

7. Vaguely worded questions as these encourage or permit vaguely worded answers. For example, the question *"How happy are you with the ways things are at the moment?"* is vague. What does "the way things are" mean? Respondents could interpret this phrase in any number of ways and give answers such as "OK" or "They could be better", which are equally vague and uninformative.

8. Ambiguous words and questions. For example, the question *"Do you have a car?"* is ambiguous. Is it asking about whether I own a car or have access to one? The question asks about 'you' but who is 'you', the respondent, the household, the family or the company?

9. Using jargon, technical terms, acronyms or abbreviations as these may not be familiar to respondents or may be interpreted in different ways by different people. For example, terms such as "social exclusion" or "hypertension" may be commonly used by researchers or health practitioners but may be meaningless or poorly understood by members of the public.

10. Using colloquialisms or words with alternative usage, as different respondents may interpret these in different ways. For example, the term 'dinner' has different meanings: to some it denotes a cooked meal, to others an evening meal. Vegetables can be known by different names in different parts of the country, for example a turnip in parts of Scotland and the north of England is a large root vegetable with orange flesh, whereas

in other parts of the country it is a small root vegetable with white flesh.

11. Beware of leading questions, for example *"When did you last telephone your mother?"* Asking this question without any prior knowledge about the respondent's mother assumes that a) the respondent has a mother who is alive, b) who has a telephone, and c) that the respondent is in contact with her. Similarly the following question asked of those who indicated that they had once been told they had cancer 'Which type of cancer do you have?' – assumes that the respondent currently has it, which might not be the case.

12. Not including don't know and not applicable codes. It is often useful to know that someone does not know the answer to a question, as this can be an important piece of information. It is also important to differentiate between reasons for no answer to a question, such as between those who don't know, refuse to answer or cannot answer because the question is not applicable to their circumstances.

13. Avoid proverbs or using stereotypes, especially when measuring attitudes, as such sayings can lead to unconscious agreement. It is better to get the respondent to think afresh about the issue.

14. Using loaded terms, such as free, democratic, natural, modern and so on.

6.8 Sources of measurement error in surveys

In a questionnaire a question may not measure the factor is was designed to detect. This is referred to as *measurement error*.

Traditionally these errors have been classified into two broad categories, those connected with survey questions and those connected with survey interviewers (Fowler et al, 1990).

More recently there has been a shift in emphasis, from viewing errors as being the product of either the questionnaire or the interviewer, to being related to the nature of the tasks the actors in a survey interview have to perform (Oksenberg et al, 1991). This task-focused classification is useful in helping us to understand the potential sources of measurement error as it focuses on the specific components of the question-and-answer process. The task-focused model on the other hand, would help to identify the cause of the problem by enabling the researcher to identify whether the problem is one of comprehension, processing or communication. These problems are summarised in Table 6.7.

Table 6.7 Components of measurement error

Traditional Model	Task-focused model
• Problems with survey questions that - are misunderstood - cannot be answered, either at all or accurately - respondents will not answer	• Comprehension problems resulting from: - use of vocabulary - complex sentence structure - not understanding the nature of the task and the rules about how to respond
• Problems with survey interviewers - do not read the questions as worded - probe directively - bias answers as a result of the way interviewers relate to respondents (for example, differences in ethnicity, age, social class, gender) - record answers inaccurately	• Validity problems resulting from: - respondents interpreting the same question in different ways, or - in the same way but not in the way the researcher intended
	• Processing difficulties - respondents may be unwilling or unable to retrieve the information necessary to answer the question
	• Pronunciation or communication difficulties - these may affect both interviewers and respondents

6.8.1 Interviewer error

Interviewers can be a source of error in surveys. There are different components of interviewer error:

- interviewer characteristics, such as gender, age or ethnicity (e.g. lower rates of anti-Semitism were reported by respondents who had been interviewed by someone who appeared to be Jewish than those who did not appear Jewish (Robinson and Rhode,1946 cited in Fowler and Mangione, 1990);
- interviewer expectations and attitudes; and
- interviewer behaviour such as
    - not reading the question as worded
    - directive probing
    - relating to the respondent in a way that affects his /her behaviour

- inaccurate recording of answers.

The following actions, put forward by Fowler and Mangione, 1990, can be taken by the researcher to mitigate against interviewer-error.

a. Questions must be carefully designed so that:

- as written they fully prepare respondents to provide answers;
- they mean the same thing to every respondent; and
- the kinds of answers that constitute an appropriate response to the question are communicated to all respondents.

b. Interviewers are given specific instructions or guidelines about how to conduct the interview. As a minimum, these guidelines should include:

- reading the question exactly as worded;
- probing inadequate answers non-directively;
- recording answers without interviewer discretion;
- maintaining neutral, non-judgemental relations with the respondent.

c. Interviewers need to be trained and this training needs to be on going.

d. Interviewers need to be supervised to ensure that they follow the guidance in b) above.

6.9 Evaluating survey questions and instruments

Due to the various sources of error that can occur when developing quesetionnaires it is important to evaluate the survey questions and instruments. Table 6.8 summarises the different methods available for pre-testing questionnaires, including cognitive interviewing methods.

Table 6.8 Methods of reviewing and testing questionnaires[15]

Method	Description
Focus groups / depth interviews / other flexible qualitative methods	Qualitative techniques widely used to explore the concepts, viewpoints and vocabulary used by the population which is to be sampled for a quantitative survey on a given topic. Flexible, exploratory approach not bound by a fixed questionnaire that incorporates preconceptions that may be false. Not fully replicable. Depends on judgement and intuition of qualitative researchers. Does not directly lead to questions suitable for a quantitative survey instrument.
Field rehearsal piloting	Rehearsal of the field data collection process as a whole. Researchers often mainly preoccupied with response, length of questionnaire and other operational issues not directly bearing on questions as measures. Coverage of question performance often sketchy. The research team sometimes conducts personal or telephone debriefing of interviewers after rehearsal pilot. Interviewers may sometimes be allowed to try limited rewordings of questions that appear not to work well and to report back on the effects of rewording. Debriefing may be backed by interviewers' notes and/or tape recordings of some interviews. Can capture the observations and intuitions of experienced interviewers; but typically each interviewer sees few cases. Debriefing may be dominated by reports on a few examples, which prove to be atypical.
Pilot respondent debriefing	Personal debriefing of respondents after they have responded to a trial version of a questionnaire, in the field or as a 'hall test'. May be conducted by interviewers or researchers. Applied shortly after the event can capture some of respondent's impressions and thought processes in responding to questions/ completing a questionnaire. Relies on respondent to identify 'problems'. Tends to be at level of whole questionnaire because of lack of time to cover all questions in detail. May be difficult for respondent to understand purpose of re-interrogation.
Dynamic piloting	An intensive, informal, small scale, iterative process. Initial version of a questionnaire is tried out by a small interactive research team on a small sample of respondents from the population. Wording or other problems are identified, changes are rapidly made and a revised version is again tested. Very time-effective. Question designers interact directly with respondents. Leads directly to a final questionnaire. Results and decisions, other than the final questionnaire, often not recorded. Based on small convenience samples. Method not fully replicable.
Split panel comparisons	Experiment comparing 2 or more set versions of quantitative question(s) or questionnaire. Allows statistical comparison of results. Experiment needs to be on a large scale for results to be useful. Sponsors may reject because only a fraction of the sample will be asked the 'best' question. Method may reveal differences in response distributions, but not reasons or which question is 'better'.
Interview re-interview	Some time after initial interview using test questionnaire respondents are re-interviewed. In the 'with feedback' version the re-interviewer has access to original response and can probe discrepancies. Good in principle for assessing reliability of response, but tends to be contaminated by recall of first interview. Respondents may misunderstand and think they have given 'wrong' answers at first interview.

[15] For a more detailed review of techniques, see Esposito and Rothgeb (1997).

Method	Description
Cognitive interviewing	Term includes a number of different techniques, including 'think aloud', card sorting etc. Based on theory of the question comprehension and answering processes. Can be used to explore and delineate respondent's conceptual space and/or to study understanding of the question, recall processes, formulation of response, internal review and articulation of response, etc.
Behaviour coding	Requires sound and preferably also video recording of test interviews. Utterances and exchanges between interviewer and respondent are coded according to a strict protocol through observation of the interview and/or review of the recordings. Questions can be scored according to the number of 'symptoms of difficulty' (e.g. request for repeat or clarification of question, pauses, expressions of uncertainty etc). Very time-consuming and laborious to carry out.
Expert review	Requires a small panel of 'experts' (usually researchers with appropriate experience) to critically review draft questionnaires and make comments and suggestions.

(Extract from: Thomas, R., Collins D., Rowlands, O. (2000) Question Testing and Survey Quality. WAPOR Seminar Proceedings on Quality Criteria in Survey Research III.)

6.10 Types of survey

There are two main types of survey, those that are concerned with providing information about a *cross–section* of the population of interest at a particular point in time and those that are concerned with providing *longitudinal information* about individual sample members over time.

There are different types of cross-sectional surveys and these are described below.

Continuous surveys take place 'continuously'. Fieldwork takes place in each month of the year, with the sample in any one month being broadly representative of the target population. Such surveys are designed to measure net annual change at the aggregate rather than individual level. Thus there is nothing in the design of such surveys that requires an overlap in sample units at different points in time. Examples of continuous surveys are the National Travel Survey, Family Resources Survey and Health Survey for England.

Repeat surveys take place at scheduled regular points in time, such as every year or every two years. Fieldwork is concentrated into a few months. Such surveys enable net change at the aggregate level to be measured, as estimates from one survey can be compared with another in the series. However, unlike continuous surveys, for repeat surveys is not possible to determine whether the observed change took place gradually or not. As with continuous surveys, there is nothing in the design of repeat surveys that requires an overlap of sample units at different points in time. Examples of repeat surveys are the Repeat Study of Parents' Demand for Childcare, the National Diet and Nutrition surveys, the National Adult Learning Surveys and the British Social Attitudes surveys.

Ad hoc surveys are one-off studies: there is no plan to repeat them at regular intervals. Although they may use questions used on other surveys it is important to bear in mind that if findings from an ad hoc survey are compared with another survey, differences observed may be the result of differences in the methodologies used by the two surveys, rather than indicating real change. Ad hoc surveys include Attitudes Towards and Experiences of Disability in Britain Survey, the National Study of the Lottery and other Gambling Activities and the Temporary Employment Survey.

Rotating panel surveys are scheduled to take place at regular intervals, or continuously, and include rotating panels; that is, people are introduced into the survey, surveyed a number of times, and then rotated out of the survey. There is no attempt to follow respondents or sample units that move or link records for individuals or sample units over time to make longitudinal estimates. Rotating panel survey designs are used where estimates of change are required to be accurate for small time periods, such as three-month periods as panel designs reduce the variance of estimates of change compared with a system of independent samples over a given time period. Furthermore, such designs enable the identification of gross change between groups or states, which may be masked by aggregate net change data because all changes at the micro level cancel each other out. An example of a rotating panel study is the Labour Force Survey.

Longitudinal studies are concerned with measuring change at the individual level.

In **longitudinal studies without rotation,** individuals or sample units are followed over time, to create a longitudinal record. Analysis is at the individual level. Such data, over time, are not suitable for generalising to the wider population. Examples of such longitudinal studies are the birth cohorts, such as the National Child Development Survey, and the English Longitudinal Study of Ageing.

Longitudinal studies can be designed to include **rotation**, which means that they follow a particular group for a specified period and introduce new sample units at specified periods, to create a longitudinal record. Data can be analysed longitudinally but also each data collection period, including new sample units, can be analysed cross-sectionally, as the study sample remains representative of the survey target population. An example of a longitudinal study with this design is the Families and Children Survey (FACS).

6.11 Further sources of information

A great deal of information about specific surveys, such as the General Household Survey or the British Household Panel Survey is available over the Internet.

What follows is a summary of some of the key Internet sites that contain useful information on government surveys and other important studies.

GHS reports can be accessed via http://www.statistics.gov.uk/statbase/Product.asp?vlnk=5756

Other National Statistics reports relating to social and welfare topics, such as the FRS, and EFS can be accessed via http://www.statistics.gov.uk/onlineproducts/default.asp - social

By scrolling up and down this page you can gain access to a range of other reports on different topics such as transport, employment and health.

For information on the British Crime Survey visit: http://www.homeoffice.gov.uk/rds/bcs1.html

For information on the Health Survey for England visit: http://www.dh.gov.uk/en/Publicationsandstatistics/PublishedSurvey/HealthSurveyForEngland/index.htm

For information on the British Household Panel Survey visit: http://www.iser.essex.ac.uk/ulsc/bhps/

For information about Birth Cohort Studies visit: http://www.cls.ioe.ac.uk/

For more information about the Families and Children Survey go to: http://www.dwp.gov.uk/asd/asd5/facs/

To find out more about the ESRC data archive go to: http://www.data-archive.ac.uk/

To find out more about the Question Bank go to: http://qb.soc.surrey.ac.uk/

To find out more about the Scottish Household Survey go to: http://www.scotland.gov.uk/Topics/Statistics/16002

6.12 Further reading

Anderson J. (1983) *The Architecture of Cognition*. Cambridge, MA, Harvard University Press.

Bailar B.B. (1989) Information Needs, Surveys and Measurement Errors. In Kasprzyk D., Duncan G., Kalton G., Singh M.P. (eds) *Panel Surveys*. New York, John Willey & Sons: 1-24.

Bechhofer F., & Paterson L. (2000) *Principles of Research Design in the Social Sciences*. London, Routledge.

Brannen J. (ed) (1992) *Mixing Methods: Qualitative and Quantitative Research*. Aldershot, Gower.

Cantril, H., & Fried, E. (1944) The meaning of questions. In Cantril H., *Gauging Public Opinion*. Princeton, NJ, Princeton University Press.

Centre for Longitudinal Studies Research Review 1999.

Clark H.H., & Schober M.F. (1992) Asking questions and influencing answers. In Tanur JM (ed.) *Questions About Questions: Inquiries into the cognitive bases of surveys*. New York, Russell Sage Foundation; 15-48.

Collins D., & White A. (1995) Making the next Census form more respondent-friendly, *Survey Methodology Bulletin*, 37: 8-14.

Converse J., & Presser S. (1986) *Survey Questions: Handicrafting the Standardized Questionnaire*. Newbury Park, Sage: 63.

Czaja R. (1998) Questionnaire testing comes of age, *Marketing Bulletin* 9: 52-66.

Czaja R., & Blair J. (1996) *Designing Surveys: a guide to decisions and procedures*. Thousand Oaks, CA, Pine Forge Press.

DeVellis, R.F. (1991) *Scale development: theory and applications*. Thousand Oaks, CA, Sage.

DeLamater J. (1982) Response-effects of question content. In Dijkstra W., & van der Zouwen J. (eds) *Response Behaviour in the Survey-Interview*. London, Academic Press.

de Leeuw E.D., & van der Zouwen J. (1988) Data quality in telephone and face-to-face surveys: a comparative meta-analysis. In Groves R.M., Biemer P.N., Lyberg L.E., Massey J.T., Nichols II W.L., & Waksberg J. (eds) *Telephone survey methodology*. New York, John Wiley & Sons; 267-286.

Dillman, D. (1999) *Mail and Internet Surveys: the tailored design method*. New York, Wiley & Sons Inc.

Dillman, D. (1978) *Mail and Telephone Surveys: the total design method*. New York, John Wiley & Sons.

Edwards P, Roberts I, Clarke M, DiGuiseppi C, Pratap, S Wentz R, & Kwan I. (2002) Increasing response rates to postal questionnaires: systematic review, *British Medical Journal*, May 2002, 324: 1183.

Flately F. (2001) The Internet as a mode of data collection in government social surveys: issues and investigations, *Survey Methodology Bulletin*, 49: 1-10.

Foddy W. (1993) *Constructing Questions for Interviews and Questionnaires: Theory and practice in social research*. Cambridge, Cambridge University Press.

Forsyth B.H., & Lessler J.T. (1991) Cognitive laboratory methods: A taxonomy. In Biemer P.P., Groves R.M., Lyberg L.E., Mathiowetz N.A., & Sudman S (eds) *Measurement Errors in Surveys*. New York, Wiley: 393-418.

Fowler Jr F.J. (1995) *Improving Survey Questions*. Thousand Oaks, CA, Sage.

Fowler Jr F.J. (2002) *Survey Research Methods: third edition*. Applied Social Research Methods Series, 1.Thousand Oaks, CA, Sage.

Fowler Jr. F.J., & Mangione, T.W. (1990) *Standardized Survey Interviewing: minimising interviewer-related error*. Applied Social Research Methods Series: 18. Newbury Park, CA, Sage.

Gribble, J.N., Miller H.G., Rogers S.M., & Turner C. (1999) Interview Mode and Measurement of Sexual Behaviours: Methodological Issues, *The Journal of Sex Research*, 36 (1): 16-24.

Groves R.M., Biemer P.N., Lyberg L.E., Massey J.T., Nichols II W.L., & Waksberg J. (eds) *Telephone survey methodology*. New York, John Wiley & Sons.

Groves R.M., & Kahn R.L. (1979) *Surveys by Telephone: a national comparison with personal interviews*. Orlando, FL, Academic Press.

Hakim C. (2000) *Research Design: successful designs for social and economic research*, Second Edition. London, Routledge.

Jenkins C., & Dillman D. (1997) Towards a theory of self-administered questionnaire design. In Lyberg L., Biemer P., Collins M., de Leeuw E., Dippo C., Schwarz N., & Trewin D (eds) *Survey Measurement and Process Quality*. New York, John Wiley & Sons Inc.: 165-196.

Knight I (1994) Changes in the sample design for the Labour Force Survey, *Survey Methodology Bulletin*, 34: 25-27.

Krosnick J.A. (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys, *Applied Cognitive Psychology*, 5: 213-236.

Krosnick J.A., & Alwin D.F. (1987) An evaluation of a cognitive theory of response-order effects in survey measurement, *Public Opinion Quarterly*, 51: 201-219.

Martin J., & Manners T. (1995) Computer assisted personal interviewing in survey research. In, Lee R.M. (ed) *Information Technology for the Social Scientist.* London, UCL Press: 52-71.

Mathiowetz, N.A., & McGonagle K.A. (2000) An assessment of the current state of dependent interviewing in household surveys, *Journal of Official Statistics*, 16 (4): 401-418.

Nicholls II, W.L., & Baker R.P., Martin, J. (1997) The effect of new data collection technologies on survey data quality. In Lyberg, L., Biemer P., Collins M., de Leeuw E., Dippo C., Schwarz N., & Trewin, T. (eds) *Survey Measurement and Process Quality*. New York, John Wiley & Sons Inc.: 221-248.

Oksenberg L., Cannell C, & Kalton G. (1991) New strategies for pretesting survey questions, *Journal of Official Statistics*, 7 (3): 349-365.

Oppenheim, A.N. (1996) *Questionnaire Design, Interviewing and Attitude Measurement*. New Edition. London, Printer Publishers Ltd.

Patton, M. Q. (2002). *Qualitative Research and Evaluation Methods*. Third Edition. Thousand Oaks, CA, Sage Publications.

Payne S.L. (1951) *The Art of Asking Questions*. Princeton, Princeton University Press.

Ritchie J. (2003) The application of qualitative research methods to social research. In Ritchie J., & Lewis J. (eds) *Qualitative Research Practice: a guide for social science students and researchers*. London, Sage Publications: 24-46.

Schuman H., & Presser S. (1981) *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context*. New York, Academic Press.

Schwarz, N., & Hippler, H. (1991) Response alternatives: the impact of their choice and presentation order. In Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., & Sudman, S. (eds) *Measurement Errors in Surveys*. New York, John Wiley and Sons, Inc. 41-56.

Sudman S., & Bradburn N.M, (1982) *Asking Questions: A practical guide to questionnaire design*. San Francisco, Jossey-Bass.

Sudman S., Bradburn N.M., & Schwarz N. (1996) *Thinking About Answers: the application of cognitive processes to survey methodology*. San Francisco, Jossey-Bass.

Thomas R., & Lynn P (eds) (forthcoming) *Survey Research in Practice.* 2nd Edition. London, Sage.

Tourangeau R. (1984) Cognitive sciences and survey methods. In Jabine T, Staf M, Tanur J, & Tourangeau R (eds) *Cognitive Aspects of Survey Methodology: Building a bridge between the disciplines*. Washington DC, National Academy Press: 73-100.

Tourangeau R., Rips L., & Rasinski K. (2000) *The Psychology of Survey Response.* Cambridge, Cambridge University Press.

Tourangeau R., & Smith T.W. (1996). Asking sensitive questions, *Public Opinion Quarterly*, 60: 275-304.

Tulving E., & Thompson D.M. (1973) Encoding specificity and retrieval processes in episodic memory, *Psychological Review*, 80: 352-373.

Turner, C.F. (1984) Why Do Surveys Disagree? Some Preliminary Hypotheses and Some Disagreeable Examples. In Turner, C.F. and & Martin, E. (eds) *Surveying Subjective Phenomena*, 2. New York, Russell Sage.

Willis G.B. (1994) *Cognitive Interviewing and Questionnaire Development: A training manual*. Hyattsville, MD, National Centre for Health Statistics.

Willis G.B., Brittingham A., Lee L., Tourangeau R., & Ching P. (1999) *Response errors in surveys of children's immunizations*. Vital and Health Statistics; 6 (8). Hyattsville, M.D, National Centre for Health Statistics.

Willis GB, deMaio T, Harris-Kojetin B. (1999) Is the bandwagon headed to the methodological promised land? Evaluating the validity of cognitive interviewing techniques. In Sirken MG, Herrmann DJ, Schechter S, Schwarz N, Tanur JM, & Tourangeau R (eds) *Cognition and Survey Research.* New York, Wiley: 133-153.

The Magenta Book: guidance Notes for Policy Evaluation and Analysis

Background paper 7: why do social experiments?
Experiments and quasi-experiments for evaluating government policies and programmes

Published: April 2005
Updated October 2007

Government Social Research Unit

HM Treasury
1 Horse Guards Road
London SW1A 2HQ

7 Why do social experiments?

7.1 Introduction

One of the most pressing problems facing evaluators and policy makers is to determine whether a policy or programme has caused change to occur in the outcomes it was designed to influence, and whether any such change has occurred in the desired direction. For example, have policies that aim to reduce unemployment led to a fall in the numbers out of work, or are some other factors responsible? Social experiments help policy makers answer these important questions.

Social experiments essentially test whether a programme or policy has led to change in the outcomes that the programme or policy was designed to affect, over and above that which would have occurred in the absence of the programme or policy. Social experiments do this by providing *potentially unbiased estimates* of the programme or policy's impact. That is, an estimate of impact that is entirely attributable to the programme or policy itself, rather than some other factor(s). For example, a new policy might seek to reduce re-conviction rates among offenders. A social experiment aims to show how much of any observed drop in re-convictions is attributable to the policy alone, rather than to some other factor(s).

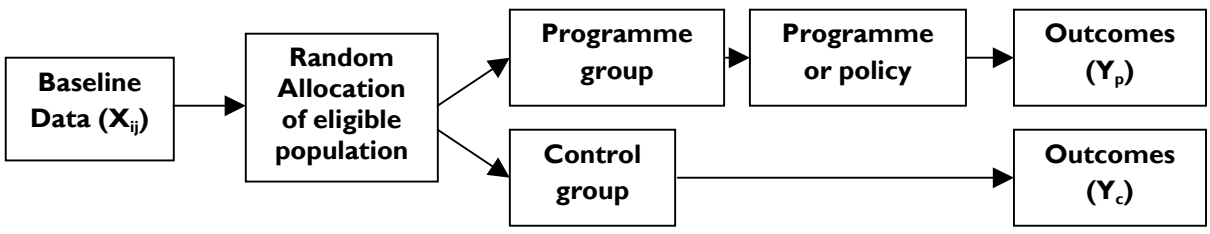
This chapter considers what social experiments are. It looks at the experimental method and its key feature, random allocation of study units to programme and control groups. Objections and limitations of random allocation are discussed, specifically ethical and analytical considerations. The second section of this chapter considers quasi-experimental designs for policy impact evaluation.

Section 1: Experimental approaches to measuring policy and programme effects

7.2 Random allocation

Central to a social experiment is the concept of random allocation, alternatively referred to as random assignment or randomisation. To understand how this works, consider the diagram below Figure 7.1, which depicts a simple random allocation design.

Figure 7.1 Random Allocation Design



A new programme or policy has an intended target population. Units in the study's target population can consist of:

- Individuals
- Institutions, for example whole schools or hospitals can be randomly allocated; or
- Whole areas, such as postcodes or wards

Prior to randomisation, it is common practice to collect baseline data on each unit in the study. Subsequent to this, units are allocated, in this example to two groups[16] at random - a programme and control group. Units allocated to the programme group go on to receive the *new* programme or policy, or be subject to a *change in an existing* policy; units in the control group are subject either to an *existing* programme or policy, or *no* programme or policy at all[17].

Randomisation is very important because, provided that the sample is large enough, it ensures that there are no systematic differences, on average, between units in the programme and control groups, at the point when they are allocated or assigned. In other words, there is no systematic relationship between membership of the programme or control groups, and the observed and/or unobserved characteristics of the units in the study. This means that any statistically significant difference in the average value of outcomes for the programme group and the average value of those same outcomes in the control group (represented as $\Delta = Y_p - Y_c$ above), measured after the new policy or programme has been introduced, result from the impact of the programme or policy alone. These impacts are statistically speaking unbiased and internally valid[18] estimates of the programme or policy's impact, given certain assumptions – we shall discuss these assumptions below.

Very often the control group is said to be an estimate of the *counterfactual*. The counterfactual represents what would have happened to the programme group had the units allocated to it not been subject to the new policy or programme, or subject to an alternative policy or programme. Many evaluators consider a control group formed at random to be the best representation of the counterfactual (Boruch, 1997, Shadish, Cook and Campbell, 2001; Orr 1999). This is because on average, units in the programme group are statistically equivalent to

[16] It is possible to assign individuals to more than one programme group.

[17] This is something on a simplification. In fact to overcome some of the ethical problems of social experiments (see section 7.5.3), some experiments involve changing a policy for all experiment participants but giving the treatment group a larger dose of the policy.

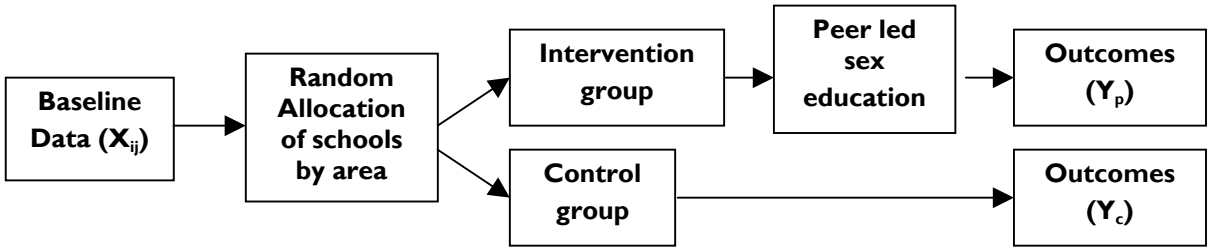
[18] There are four main types of validity associated with experimentation: statistical conclusion validity, internal validity, construct validity and external validity. Social experiments have been shown here to possess internal validity when properly implemented. In order to determine whether a social experiment possess these other forms of validity, additional features of an experiments design need to be considered. For a detailed discussion of validity within the context of social experimentation see Chapter 2 in Cook and Campbell (1979). Economists often refer to unbiased or 'internally valid' estimates of a programme's impact as being free from selection bias (see Burtless 1995).

units in the control group, except for the fact that the latter are exposed to the new programme or policy being tested.

The issues raised above are perhaps best illustrated with an example. A good example of a randomised controlled trial was the RIPPLE evaluation (see Stephenson et al, 2003). This was a randomised trial of peer-led sex education in schools in England. The aim of the evaluation was to achieve a better understanding of how a particular sex education strategy, the peer led method of teaching and learning, impacted on the knowledge, attitudes and behaviour of young people.

The study involved comparing schools that received the peer led sex education programme with schools that did not. The peer led programme was implemented in half of the schools (the experimental schools) while the remaining schools (the control schools) continued with their usual sex education curriculum.

Figure 7.2 The RIPPLE Random Allocation Design



It is important to note that random allocation to the control condition did not mean that children in this group received no treatment. The children continued to receive whatever sex education they would normally have received. This is sometimes known as the 'treatment as usual' condition. Most evaluations take this approach. They compare an experimental treatment with 'treatment as usual'. The 'treatment as usual' group thereby acts as the counterfactual.

The schools in the RIPPLE study were randomly allocated to either control or experimental status. The aim of the randomisation was to make the two groups equivalent in all respects apart from the intervention (the peer led sex education). Therefore any observed difference on the outcome measure should have been the result of the intervention and the intervention alone, because this should have been the only difference between the two groups. All other sources of bias should have been controlled by the random allocation process. This appeared to be the case, at least in respect of a small number of variables that were measured following the random allocation (see Table 7.1).

Table 7.1 The effect of randomisation in creating equivalent control and experimental groups

%	Completed questionnaire	Free school meals	5+ GCSE (1997)	Privately owned housing	White	Dislike school	Had sex
Control	91 (n=4250)	11.1	46.6	30.6	92.0	19.4	6.7
Experimental	93 (n=4516)	10.0	46.8	26.3	90.2	19.4	6.7

It should be noted that it is important to collect sufficiently rich data in any form of evaluation, even random experiments. This allows one to determine whether the allocation process has distributed cases randomly, across variables known to have an important influence on the outcome measure.

7.3 Assumptions underpinning social experiments

For social experiments to provide unbiased estimates of programme impacts a number of conditions should hold. Estimates must possess both internal and external validity. Even when some of these conditions do not hold, however, results from social experiments still provide policy makers with useful findings. Furthermore, the statistical power of the experiment must be sufficient to be able to detect programme impacts should they exist. The concept of statistical power and the Minimum Detectable Effect are outlined at Box 7.1 in the Appendix at the end of this chapter; statistical power and statistical error in experimental design are explained at Box 7.2.

7.3.1 Internal validity

Internal validity refers to the most robust findings that result from a social experiment being properly implemented. There are a number of assumptions that must hold for impact estimates from social experiments to be considered 'internally valid'. In this section a number of commonly occurring 'threats' to internal validity within the context of a social experiment are considered. These 'threats' are in addition to the standard threats to validity common to all forms of social research, such as general Hawthorne Effects. There is however the possibility that Hawthorne Effects unique to social experimentation may exist, though there is no evidence that empirically verifies their existence.

Scriven (1991: 186) defines Hawthorne Effects as:

> *"the tendency of a person or group being investigated, or experimented on, or evaluated, to react positively or negatively to the fact that they are being investigated/evaluated, and hence to perform better (or worse) than they would in the absence of the investigation, thereby making it difficult to identify any effects due to the treatment itself".*

As a result, where individuals or institutions are knowingly assigned at random to either programme or control groups, the fact that units are aware of their control (or experimental) status may alter their behaviour. For example, if units who opt to enter a trial training programme are subsequently allocated to a control group, they may seek to gain access to the same or similar training by some other route, though they probably would not have done so had they not been randomly allocated. In such circumstances the control group no longer represents the counterfactual of 'no training', and if such training produces an impact among members of the control group who receive it, impact estimates generated through a comparison between programme and control groups members will be attenuated and biased. Cook and Campbell (1979) refer to such an occurrence as 'compensate rivalry'; while Heckman and Smith (1995) discuss a similar phenomenon they refer to as 'substitution bias'.

In medical trials, mechanisms such as 'blinding' and concealment are used in an attempt to ensure that 'compensated rivalry', and other Hawthorne type effects, such as 'performance bias' and 'resentful demoralisation' (factors discussed below), do not occur and thereby confound impact estimates. Clark and Oxman (1999: 3) define blinding as:

> *"Keeping secret group assignment (e.g. to treatment or control) from the study participants or investigators. Blinding is used to protect against the possibility that knowledge of assignment may affect patient response to treatment, provider behaviours or outcome assessment."*

Blinding is seldom, if ever, possible in social experiments.

It is also possible that programme administrators may be tempted to provide members of the control group with access to services similar to those being received by programme group members, or seek to improve services for which control group members are eligible. Such a phenomenon is referred to as 'performance bias' (Clarke and Oxman (eds.), 1999). Cook and Campbell (1979) also identify a problem of 'resentful demoralisation', whereby those allocated to the control group, in the knowledge that they are not receiving services available to the programme group, are discouraged and perform worse that they would have in the absence of the research. If this occurs, impact estimates will be biased and inflated above their true value.

In addition, for 'internal validity' to hold, programme and control groups need to be statistically equivalent not just at the point in time they are randomly allocated, but also at the point in time that outcomes are measured. During the period between random allocation and the measurement of outcomes, units may be lost from the experiment in ways that are systematically related to their random allocation status and, therefore, the processes of attrition from the experimental samples may differ across programme and control groups. As mentioned earlier it is very important that sufficient data is collected, across a range of

relevant variables, to allow the evaluator to know what kinds of individuals are dropping out/attriting from each group.

The problem of 'attrition' from the study is particularly prominent where surveys are used to measure programme outcomes. If differential rates of attrition across programme and control groups occur, the internal validity of experimental impacts would be compromised. Steps would need to be taken to adjust programme impact estimates, for example, through the application of non-response weights to the data or through the use of quasi-experimental impact estimation methods. Such approaches to recovering unbiased estimates of programme impacts are referred to as 'intention to treat' analysis in the clinical research literature[19]. Levels of attrition will also have implications for the external validity of impact estimates.

Finally, two phenomena termed 'crossovers' and 'contamination' also affect the internal validity of impact estimates. Crossovers occur where there is a fault in the process of random allocation and individuals who should be assigned to the control group, are assigned instead to the programme group, or vice versa. Such a fault could result either from allocation being non-random in some way, or through deliberate subversion on behalf of programme or administrative staff responsible for randomisation. The end result is that statistical equivalence between programme and control groups is violated and impact estimates are likely to be biased.

Contamination refers to the situation where individuals assigned to the control group at random, actually receive services through the policy or programme being tested – that is they receive services set aside for the programme group. Such a situation can arise either as a result of administrative error or deliberate subversion. Where programme services are received by control group units in error and those services have an impact on outcomes, impact estimates will be biased downwards.

7.3.2 External validity

External validity refers to the degree to which 'a causal relationship holds over variations in persons, settings, treatments, and outcomes' (Shadish, Cook and Campbell, 2002: 21). Alternatively, evaluators may wish to be able to infer causality from a particular social experiment to the population from which the units in the particular study were drawn. In lay terms, external validity refers to whether or not the findings of a social experiment are reproducible in 'real world' contexts.

Social experiments are often implemented in the form of a policy demonstration (or pilot test), run in a limited number of geographical

[19] Intention to treat analysis is a procedure whereby the data on all participants in a trial are included and analysed in the arm of the trial to which they were allocated at random. This is so whether or not the participants actually received or completed the intervention given to that arm. It is important to do this in order to avoid the bias that is derived from the loss of participants to the trial.

areas, on a sample of the target population. The results from these pilots are often used to determine the implementation of the policy in different settings, contexts and time periods, or for the target population as a whole. Before going on to consider some of the threats to achieving external validity, it is worth pointing out that quasi-experimental pilots, prototypes or pathfinders are also prone to difficulties relating to generalisability. Furthermore, as Orr (1999: 14) states:

> *"True external validity…… is an ideal that is almost impossible to attain, if only because the continually evolving policy process represents such a moving target. Nevertheless, it is an important ideal to strive for, and in assessing the strengths and weaknesses of alternative evaluation methods or results, it is important to gauge their external validity as well as their internal validity".*

Donald T Campbell, a major advocate of randomised controlled trial as a means of ensuring internal validity, also acknowledged the importance of external validity. He referred to external validity as 'situation specific wisdom', and argued that without it researchers will be 'incompetent estimators of programme impacts, turning out conclusions that are not only wrong, but are often wrong in socially destructive ways (Campbell 1984:42).

7.3.3 Threats to external validity

Policy context

As this above statement from Orr suggests, one of the major threats to the external validity of results from social experiments is the speed at which the concerns of policy makers change. By the time results emerge, policy considerations may have moved on and the environment in which policy is made substantially altered. Weiss (1998) argues that in cases where the policy environment is subject to change and the rapid development of new ideas, as is often the case in areas where policy development is at an embryonic stage, it is advisable to consider alternative methods to evaluate policies or programmes or alternatively to ensure that the existing evidence on likely policy or programme impacts is harnessed to its full potential. However, it could also be said that the external validity of these alternative methods is also threatened in a changing policy environment, if data is being collected over time.

Site selection

One crucial element in determining the external validity of a social experiment is the method used to select the areas or institutions where the social experiment is implemented. Ideally, the evaluator would draw up a full list of all areas/institutions and select a large sample of them on a random basis. Then within each area or institution, the eligible population at which the programme or policy is targeted is allocated at random to programme and control groups. Alternatively, and if

possible, the evaluator might select individuals across the jurisdiction as a whole (such as the whole of Great Britain as example) to be allocated at random. The objective of such an approach is for the probability of inclusion in the study to be known for all units, where the area or institution acts as a primary sampling unit. In reality, such an approach is seldom possible and indeed, selecting a large number of sites at random tends to be too expensive. Moreover, there may be resistance in some selected areas or institutions to running a social experiment. As a result, areas or sites are often selected as either *convenience* samples, whereby sites are included because they were easily recruited; or selected *purposively*; whereby the sites are selected because they are well matched to the population of interest in observable characteristics. In such cases the latter is preferable to the former (Orr 1999).

Randomisation bias

Some authors, notably Heckman and Smith (1995), argue that the existence of a social experiment may alter the composition of the sample of individuals entering the programme at a given site or area. In other words, individuals may refuse to be randomly assigned and therefore the sample of individuals entering the experiment might not be representative of the target population at that site. Heckman and Smith (1995) refer to this phenomenon as 'randomisation bias'.

Scale bias

The external validity of social experiments can also be compromised by what is known as 'scale bias' (see Manski and Garfinkel 1992 and Garfinkel, Manski, and Michalopoulos 1992). Scale bias refers to the fact that social experiments run as demonstrations may fail to capture community wide effects that occur when the policy or programme is introduced universally. Patterns of community or 'macro' behaviour unaffected by a demonstration or limited social experiment may come to the fore when the entire target population are exposed to the programme or policy.

Substitution and displacement effects

If one imagines a welfare-to-work programme where individuals are provided with help from a caseworker to return to work, it is possible that members of the programme group might obtain jobs at the expense of those outside the experiment or those in the control group. These effects are known as *substitution* effects. In an experiment to measure the impact of a programme to reduce domestic burglary, where the programme is successful, substitution effects might occur where other forms of crime, for example vehicle theft rise. Moreover, a successful experiment operating in one pilot area may disperse criminal activity to nearby areas where the programme under investigation is not operating. Similar effects identified in studies of active labour market programmes are referred to as *displacement effects*. The seriousness or otherwise of such an effect is dependent on the context in which the demonstration

is operating and in the case of labour market programmes, the tightness or otherwise of the labour market. A range of methods are available in order to explore the context in which programmes are operating, including qualitative and programme theory approaches to evaluation.

7.4 Advantages of randomised trials

Randomisation in theory provides a single or series of impact measures which, if the experiment is properly designed and its control requirements adhered to, provide strong evidence of programme impacts net of confounding factors. The additional advantages of randomisation can be summarised as follows:

- As illustrated in Example 7.1 (see below) results from social experiments are clear and easy to explain policy makers, Ministers and other non-technical audiences (Burtless 1995 and Orr 1999);
- Other methods that seek to answer causal questions concerning policies or programmes, known as quasi-experimental methods, require the application of *complex* statistical techniques based on assumptions that in many cases are unlikely to hold. This is not the case with randomised controlled trials;
- Randomisation can be a fair mechanism of allocating interventions where interventions are dependent on limited resources and need therefore to be rationed;
- Impact estimates from social experiments, combined with measures of the costs of the policy or programme being evaluated, can easily be incorporated into a formal cost-benefit analysis. The measurement of benefits or programme impacts as well as a programme's net costs are relatively straightforward in an experimental setting (see Boardman, Greenberg, Vining and Weimer, 2001).

Results from various social experiments attempting to measure the impact of similar policies or programmes can be combined using methods of research synthesis and meta-analysis. Techniques for combining results from a number of social experiments are well-established (Cooper and Hedges 1994, Lipsey and Wilson 2001). Combining studies in this way adds to the confidence in determining whether programmes or policies are effective across a variety of settings and contexts.

Example 7.1 Clarity of Results from Social Experiments

As an example of the simplicity with which experimental results can be presented, Table 1 displays estimated impacts from a social experiment carried out in California known as GAIN. GAIN's objective was to provide a range of services to individuals who were out of work and claiming welfare in order to help them obtain a job. Individuals eligible for the programme were allocated at random to either a programme group or a control group. Only individuals assigned to the programme group could receive the services provided through the GAIN programme.

Example of Results from a Social Experiment

All participating GAIN counties	Programme Group Members	Control Group Members	Difference
Ever employed (%)[1]	56.7	50.8	5.9***
Average total earnings ($)[2]	7,781	6,367	1,414***
Sample size (N=)	17,677	5,114	

Notes:
(1) Individuals having spent some time in a job over a three year period subsequent to random allocation.
(2) Total average earnings for individuals over a three year period including those in the sample whose earnings were zero.
*** indicates statistical significance at the 1 per cent level.
Source: adapted from Table 4.1 (Riccio, Friedlander and Freeman, 1994: 122)

The table above shows that the GAIN programme produced positive effects in both employment and earnings for those assigned to the GAIN programme group, over and above rates of employment and levels of earnings in the control group. For example, the table shows that 56.7 per cent of individuals assigned to receive GAIN services were in work of some form over a three-year period. This compares to 50.8 per cent of those assigned to the control group. The impact of being invited to receive GAIN services on working was a 5.9 percentage point improvement in rate of employment (56.7 minus 50.8). This represents a 12 per cent increase in employment over the three years post-random assignment – an increase solely attributable to GAIN.

7.5 Disadvantages of randomisation

The randomised controlled trial (RCT) has impressive advantages as a method of evaluation, but it has some disadvantages. The criticisms and difficulties associated with RCTs can be grouped under four headings (see Burtless and Orr 1986): policy utility, methodological, ethical/practical and cost.

7.5.1 Policy utility

Commentators have criticised RCTs on the grounds that they do not address many of the questions of interest to policymakers (see Heckman and Smith 1995, and Pawson and Tilley 1997, for examples of critiques of RCTs addressing this issue from different perspectives). In many of its guises, critiques of this nature refer to what is known as the 'black box'

problem. An RCT 'produces a description of outcomes, rather than explanations of why programmes work' (Pawson and Tilly 1997: 30). Clearly policymakers are interested in what it is about their policy or intervention which leads to change. Questions of this nature are important for those responsible for implementation and delivery. It though is also worth saying that these arguments are not unique to RCTs and could also be applied to quasi-experiments.

Heckman and Smith (1995: 95) suggest that in most practical cases, RCTs will not be able to answer important questions such as which factors affect the decision of individuals to take part in non-mandatory interventions? What is the importance of area effects on results? What influences decisions to drop out of programmes? What are the costs of various interventions? 'Some of these questions might in principle be evaluated using random assignment designs, but practical difficulties would make it impossible in most cases' (Heckman and Smith 1995: 95).

Whilst RCTs usually cannot inform many black box problems and implementation issues, they do provide valid information on the likely impact of a policy or programme and the variation in effect sizes across different research sites. The reasons for these variations in effect size, and other implementation issues, are usually best addressed using qualitative, consultative and other formative methods of evaluation. However, it is sometimes possible to test implementation issues and some black box processes by designing a structured experiment, that evaluate a number of delivery mechanisms.

7.5.2 Methodological

There are a number of methodological criticisms of RCTs. Some of these have already been discussed, particularly in relation to external validity and Hawthorne Effects, but some additional problems are noted below. However, it is worth repeating that these issues are not unique to RCTs and apply equally to quasi-experimental approaches.

- RCT evaluations are often unable to detect, or simply do not look for, *unintended* consequences of a policy or programme (Heckman and Smith 1995: 99).
- Where data on outcomes (post-test) are collected using a survey, members of the control or treatment groups might refuse to take part in the survey leading to missing information about outcome variables (Burtless and Orr 1986: 615). If the rate of non-response across treatment and control groups is non-randomly distributed, the simple gain score estimate above struggles to detect an unbiased treatment effect (Blundell and Cost Dias 2000:434). Units allocated to receive a treatment may also, for one reason or another, not go on to receive treatment. If this problem is widespread the analysis of data from the experiment using a simple gain score can lack

the necessary statistical power. Gain score estimates under these circumstances will also be biased.

- RCTs generally provide average impact estimates or gain scores. An average score can hide variations in treatment impact across those who receive treatment. This is especially problematic where treatment effects are heterogeneous, that is having a large variance. For example, results from an RCT evaluation might show that on average the treatment group does well on a post-test score, but that this positive average score masks a negative impact among a subgroup (see Davies, Nutley and Tilley 2000: 263). This issue though could be addressed if one has collected relevant data for both the experimental and control groups. This would provide for comparative sub-group analysis. Furthermore, where an RCT is undertaken across a number of sites (e.g. different labour market areas) it is usually possible to detect important variations in effect size, thereby providing valuable information about the range of impacts a policy is likely to have.

7.5.3 Ethics

A charge commonly levelled at social experiments by administrators, policy makers and even some evaluators and analysts is that they are unethical. This is said to be because individuals allocated to the control group are barred from receiving the services available to the programme group, and thus members of the control group are being discriminated against. Such attitudes may also stem from a sense of unease concerning the ethics of 'experimenting' on human subjects.

Such a charge is countered through the assertion that prior to the results from the social experiment becoming available, the supposed scale and direction of any impact on units in the study is unknown or equipoise. In other words, policy makers and evaluators have no way of knowing in advance whether those assigned to the control group are worse off than they would have been had they been assigned to the programme group instead, or that those assigned to the programme group will be better off. Furthermore, *if* there is good reason to suspect that, a *priori*, a programme will produce benefits for those units receiving programme services, there is no evidence as to whether these benefits accrue at an unreasonable cost to society. The justification for launching a social experiment is that policy makers are unsure of whether the policy or programme generates the benefits it was designed to achieve. In some cases, once a social experiment has shown that some programme or policy produces positive effects for those in the programme, the experiment can be stopped and members of the control group given access to the services that have been proven effective.

There are, however, some circumstances in which it is unethical to mount a social experiment. These are where (Rossi, Freeman and Lipsey 1999, Orr 1999, Cook and Campbell 1979):

- members of the control group are denied services which are *known* to be beneficial on the basis of *existing evidence*;
- members of the control or programme groups are subject to phenomena *known* to have harmful effects or outcomes; and in most cases;
- Members of the control group are denied access to services to which they have an *historical entitlement*.

In order to address concerns regarding the ethics of social experimentation, participants are frequently asked to provide informed consent to randomisation (Boruch 1997, Orr 1999). Individuals have the experiment described to them in detail and are asked to provide written consent to be randomly allocated. What constitutes 'informed consent', however, is often contentious or uncertain.

7.5.4 Cost considerations

In most cases mounting a RCT is no insignificant undertaking. Put simply, social policy experiments are complex and expensive. Results from RCTs also take time to become available, in most cases at least two to three years. These points though apply to most forms of evaluation and should not be taken as arguments against RCTs. There are though some questions that evaluators should consider in deciding whether an RCT is the appropriate evaluation methodology for a particular policy or programme:

- Is the budget for my policy large enough to warrant a full-scale summative evaluation using RCT methods? You should bear in mind when making this judgement that the costs of an RCT may be more than offset by the social benefits of having the information or results they deliver available to policymakers (Burtless and Orr 1986: 626).
- Can the decisions I will need to take based on findings from my evaluation wait until results from an RCT become available?
- Does my department have the administrative capacity to deliver a complex evaluation like this?
- Does the available external analytical community have the competence and capacity to undertake an RCT to the highest possible standards?

Evaluators should note the advice of Cook and Campbell:

> *"The case for random assignment has to be made on the grounds that it is better than the available alternative for inferring cause and not on the grounds that it is perfect for inferring cause"*
>
> (Cook and Campbell 1979: 342)

It should also be noted that quasi-experiments and other types of research and evaluation can also be very expensive because these studies usually require a significant amount of data collection. These other methods also cannot usually deliver the same degree of validity and reliability of findings, free of the biasing effects of confounding factors.

Section 2: Quasi-experimental approaches to measuring policy and programme effects

Thus far the experimental method has been discussed as it applies to the evaluation of social programmes and policies. In this section we consider the group of summative evaluation methods known as quasi-experiments.

Cook and Campbell's (1979) classic text on quasi-experimental design defines a quasi-experiment as:

> *"Experiments that have treatments, outcome measures, and experimental units, but do not use random assignment to create the comparisons from which treatment-caused change is inferred. Instead, the comparisons depend on non-equivalent groups that differ from each other in many ways other than the presence of the treatment whose effects are being tested".*
>
> (Cook and Campbell 1979: 6)

Quasi-experimental methods are basically applied in situations where the degree of control over the policy or intervention required to apply random assignment is not available to the evaluator, or the application of random assignment is felt to be unethical. For example, a policy or intervention may be introduced universally to the entire eligible target population, leaving no scope for randomising-out a proportion of those eligible to form a control group.

Like social experiments, quasi-experiments seek to meet all the conditions required to infer whether a causal relationship exists between a programme or policy and a given outcome. These conditions are set out below:

1. That the presumed causal factor occurred prior in time, or precedes, the observed effects – this is achieved by experiments or quasi-experiments through manipulation of the causal factor (the policy or programme) usually in the context of a policy pilot or demonstration;

2. That the causal factor co-varies with, or is related to, the observed effects – ascertained through statistical analysis; and

3. That other extraneous or confounding factors have been ruled out as varying with the observed effect or accounting for all of the observed effect (in the case of social experiments, this final condition is met through the use of random assignment).

In such circumstances, it is the convention that the 'policy off' group, where one exists, is referred to as a *comparison* rather than control group. The use of the term control group is generally reserved for cases where access to the new programme or policy is determined through random assignment (with exception of some case control studies). As is the case with a control group in an experimental design, outcomes for a comparison group within the context of a quasi-experimental design represent an *estimate* of the counterfactual.

Non-experimental methods of evaluation are *not* covered here. By non-experimental methods, we mean estimating the impact of some 'naturally occurring' phenomena not under the control of the policymaker or evaluator, and therefore phenomena that cannot be directly manipulated, assigned in a controlled way or piloted. Non-experimental methods (sometimes referred to as *natural experiments*) have been used to determine the impact of smoking on the incidence of lung disease for example, or the impact of divorce on child outcomes. In other words, non-experimental methods are used to evaluate what Cook (2002) refers to as *non-manipulable* causes. Such studies are most effective where impacts are estimated using panel or cohort data, or through the use of matching techniques (including retrospective case-control studies) similar to those methods implemented in quasi-experimental settings. It is important to note that some clear theoretical hypotheses about the relationships under investigation are essential, otherwise it is often difficult to disentangle causal relationships.

7.6 Approaches to quasi-experimental impact estimation

There is a wide-range of quasi-experimental approaches that aim to provide valid measures of programme or policy impacts. The literature on quasi-experimental methods is large and due to the limitations of space only a few of the more commonly used or innovative quasi-experimental approaches are discussed in detail. A very useful guide to experimental and quasi-experimental designs can be found in Shadish, Cook and Campbell (2001) and in Campbell and Russo (1999).

Each of the main quasi-experimental methods listed below seeks to estimate the counterfactual by a variety of means, in contrast to the experimental method discussed above, which does so through randomisation. In essence, each of these approaches aims to establish the existence of a causal link between changes in an outcome of interest

to policy makers and the programme or policy under investigation. Quasi-experimental methods include:

- Single group pre and post-test, or before and after designs;
- Two group pre and post-test, or before and after design (alternatively referred to as non-equivalent comparison group (NECG) pre and post-test design);
- Various extensions of the two-group design that involve statistical matching – including cell matching and propensity score methods;
- Interrupted time series design;
- Regression point displacement design;
- Regression discontinuity design; and
- Other econometric methods (including the use of instrumental variables and Heckman (1979) style selection models).

Each of these methods can be used to evaluate *manipulable causes* such as government polices and programmes, particularly in the case of policy pilots or demonstrations. The regression discontinuity design, single group pre and post-test design, the non-equivalent comparison group pre and post-test design are discussed below. Readers interested in a fuller discussion of quasi-experimental designs are referred to Shadish, Cook and Campbell (2002) and Cook and Campbell (1979).

7.6.1 Single group pre and post-test design

The single group pre- and post-test design is a common means of estimating the impacts of policies or programmes. Generally such designs are considered to be weak because they are unable to account satisfactorily for a wide variety of alternative explanations for any observed programme impacts. That is, this type of design does not really provide any valid and reliable information about an independent counterfactual (i.e. what would have happened to a comparison group if the policy or programme had not been offered or if some other intervention had been provided). Indeed Campbell and Stanley (1963) use it for heuristic purposes as a means of illustrating the full range of factors that can undermine internal validity in quasi-experimental evaluation. It should be noted, that this design can be extended in a number of ways to improve the validity of programme impact estimates. For example, the Department for Work and Pensions (formerly the Department of Social Security) used an elaborated version of the single group design to evaluate the welfare benefit called 'Jobseeker's Allowance' (Smith, et al, 2000).

Example 7.2 Evaluating Jobseeker's Allowance: a Before and After Study

The Department for Social Security, Department of Education and Employment, Employment Service and Benefits Agency conducted an evaluation of Jobseeker's Allowance (JSA), in order to find out whether JSA met its objectives. Prior to the introduction of JSA, a sample of people unemployed and claiming benefit (the forerunner to JSA) were surveyed and information about the group's knowledge, attitudes and behaviour was collected. This group was re-interviewed six months after the first interview but still in the period prior to the introduction of JSA. Ten months after the introduction of JSA, another sample of those unemployed and claiming benefit was drawn and surveyed. This sample was also re-surveyed six months later (Smith et al 2000: 10).

The design for this evaluation is a type of before and after design known as a 'cohort' study. Effectively we have a pre-JSA cohort on whom two pre-tests are administered and a post-JSA cohort upon which two post-tests are performed. Having effectively two pre and post-test observations aims to counter the threat of maturation to causal inference in the study. Having two pre-test observations for example, means that evaluators can determine the trend in employment and other sample characteristics among the population subject to the policy intervention before the policy is introduced[20].

Indeed, a single group pre and post-test design as described here is seldom implemented without some additional refinements (or design controls) in order to improve the capacity to draw valid causal inferences[21]. For the purpose of explaining the design and some of its limitations, however, a simplistic variant is described here.

Figure 7.3 Single Group Pre and Post- test Design

Figure 7.3 illustrates a simple single group design. The programme or policy under investigation is directed at a target population or a subset of this population in the form of a group or cohort. Prior to the introduction of the new policy or programme, data are collected on the outcomes (or dependent variables) that the policy programme seeks to influence 'Y_{t-1}'. This stage in the design is referred to as the baseline data collection stage or pre-test.

Once baseline data have been collected, the new policy or programme, or policy change can be introduced. At some point following the introduction of the programme or policy, follow-up or post-test data is

[20] The evaluators on this particularly project also adopted a form of weighting. This involved weighting the profile of pre-test observations to the profile of respondents on the post-test with respect to distribution of the latter sample by the local level of unemployment. This approach was adopted to deal specifically with the problem of maturation in the form of the changing pattern in unemployment over the lifetime of the evaluation. For more details see Smith at al (2000: 13).

[21] For example, linear maturation threats can be countered through the addition of two pre-test or baseline data collection stages.

collected on outcomes 'Y_t'. The impact of the new policy or programme, or 'Δ', is simply computed as 'Y_t *minus* Y_{t-1}'. This computation can be adjusted (using regression analysis) to account for measured factors known to affect outcomes other than the programme (statistical controls – in contrast to design controls mentioned above). Such an adjustment attempts to control for changes in background variables that might have influenced 'Δ' independently of the effect of the programme or policy under investigation.

The problem with this design is that many rival events and factors could be responsible for 'Δ' other than the new policy or programme and changes in measurable background characteristics. These rival events or factors are often referred to as 'confounds' or 'threats' to internal validity. In order for 'Δ' to be an unbiased estimate of the impact of the programme or policy being evaluated, the assumption that no unmeasured change (that is change we can not control for in a regression model) would occur (i.e. that 'Δ' = 0) in the absence of the policy or programme must hold. As we will see, this is a very strong assumption (Campbell and Kenny 1999) that is highly unlikely to be plausible in most contexts. It is for this reason that a single pre and post-test design is considered weak in terms of internal validity. Very often, however, evaluators resort to using such designs where random allocation is not possible for either design, political or administrative reasons, where a comparison group is unavailable, or where the evaluator might wish to concentrate on achieving external validity – such as accounting for scale bias (discussed above). In the latter case, the evaluator may make a conscious decision to trade-off internal validity for external validity.

Shadish, Cook and Campbell (2002) report a typology of confounds or threats to internal validity. The interested reader should refer to this source for a fuller discussion of the threats to internal validity in experimental and quasi-experimental research, but briefly the main threats can be summarised under nine headings.

Ambiguous temporal precedence

When it is not clear which variable occurred first confusion arises between which variable is the cause and which is the effect.

Selection

Systematic differences in the characteristics between the treatment and the comparison groups (Note: this does not apply to the one group pre and post test evaluation but does apply to the two-group-pre-and post-test design- see below).

History

Events that occur concurrently with treatment could produce the treatment effect.

Maturation

Naturally occurring changes over time that could account for a treatment effect.

Regression/ regression towards the mean

When units are selected because of their extreme scores these scores can regress towards their average on re-measurement.

Attrition

Loss of respondents to treatment can produce biased results if the drop-outs are different from those who remain within the group.

Testing

Exposure to a test can result in changes to scores on re-testing that are independent of treatment.

Instrumentation

The measure may not be reliable or it may change over time and these effects could be confused with a treatment effect.

Additive and interactive effect of threats to internal validity

The impact of a threat can be added to that of another threat or may depend on the level of another threat.

The point to note is that a well-designed social experiment (using random allocation methods) deals effectively with each of these threats to internal validity. Three of the most important threats to internal validity are discussed below: history threats, maturation threats and regression to the mean.

7.6.2 History threats

A 'history' threat to causal inference occurs where some event or events, between baseline (the pre-test) and follow-up (the post-test), lead to changes in outcomes independently of the programme or policy under investigation. For example, in an evaluation to measure the effect of new classroom teaching materials, a change in teaching personnel might also occur between baseline and follow-up, thereby making it difficult to ascertain whether changes in the follow-up or post-test result from the effect of the materials or change in personnel. Where such 'history' threats exist, unless such a threat is identified, measured and controlled for in any analysis, change in outcomes might erroneously be attributed solely to the policy or programme. In most policy or programme evaluations there are numerous history events which complicate, or make extremely difficult, the assessment of programme or policy effectiveness. Introducing a comparison group of units with similar characteristics to those exposed to the policy or programme

under investigation can help to control for history effects. The extent of design control for history threats gained from adding a comparison group depends on the degree to which the comparison group is exposed to the same magnitude of history effects as the programme group.

7.6.3 Maturation threats

A 'maturation' threat occurs where some fraction of the estimated impact stems not from the influence of the programme or policy under investigation, but simply from the passage of time between baseline and follow-up. Differences between baseline and follow-up values may capture the effects of some underlying secular trend causing the outcome variables being measured to change independently of the programme or policy being evaluated.

Consider a health initiative that aims to encourage young mothers to stop smoking, and which is evaluated using a before and after design. A cohort of young mothers who smoke on average more than a certain number of cigarettes a day receive a specified intervention and some time afterward the rate of smoking among the group (defined on the same basis) is measured again. During the period between baseline and follow-up, particularly over a long period of time, the rate of smoking among mothers of this age group would be expected to fall naturally as a certain proportion of individuals give up smoking as they get older independently of the intervention under investigation. This effect or trend is a maturation effect and occurs simply because the research group concerned have got older. As a result, when interpreting 'Δ' in this case, it is important to consider how much 'Δ' would have changed in the absence of the programme as a result of maturation. Such a requirement provides a strong justification for inclusion of a comparison group comprising of similar units along side a programme group in any design.

7.6.4 Regression to the mean

Regression to the mean is a common but not particularly well-understood phenomenon, which can occur for a number of reasons. In the case of a pre-test/post-test evaluation design, regression to the mean usually arises because the programme or policy is directed at individuals possessing some extreme characteristic (for example, low income, low attainment in school tests, high rates of criminal recidivism and so on). The general pattern is that units with high pre-test scores tend to score lower at post-test, and units with low pre-test scores tend to score at higher at post-test – in other words extreme scores tend on re-test to move or regress toward the mean test value.

An intuitive way to understand this is to consider the evaluation of extra reading lessons for children aged 7. Imagine that instead of all students aged 7 entering the programme, only those known to have poor reading skills are eligible. Thus, on the basis of some pre-test or baseline

measurement of reading ability, individuals in the bottom quartile of test scores are assigned to the programme. As the measurement of ability at the pre-test stage is imperfect, and therefore statistically unreliable to some degree, a fraction of the individuals eligible for the programme would not be eligible if their true test score were known without error. These individuals will in all probability score higher at the post-test irrespective of whether they attended extra classes or not. This means that the average test-score for the group will improve regardless of whether students at whom the extra help is targeted attend extra classes or not. In many cases *extreme group selection* presents this problem – average test scores will improve or decline because of mean reverting tendencies in variables measured at two or more points in time among groups sampled from the extremities (either positive or negative) of some distribution. As a result 'Δ' may to a lesser or greater extent be the result of regression to the mean.

It is important to note that regression to the mean is a potential threat to internal validity in many evaluation designs, with the exception of random allocation and regression discontinuity. In the case of the former the effects of regression to the mean are randomly distributed between programme and control groups and do not therefore affect mean comparisons of outcomes between programme and control groups. In regression discontinuity designs, the regression line, the centre-piece of the approach, effectively controls for mean reversion (see Shadish, Cook and Campbell for further details). The interested reader should consult Campbell and Kenny (1999) for a fuller discussion of regression artefacts.

7.7 Non-equivalent comparison group designs (or two-group-pre-and post-test design)

The non-equivalent comparison group (NECG) design involves the evaluator selecting a group of units similar to those receiving the new policy or programme that is being tested. Such a group is called a comparison group (similar to a control group in a social experiment) and acts as a counterfactual. As we have seen, estimation of a counterfactual is essential to the process of causal inference or attribution. The concept underlying such selection is to obtain a comparison group that is as similar as possible to the programme group in all respects. It is a stronger form of design to the one-group pre and post test design (discussed above) because it includes a comparison group.

Figure 7.4 Non-equivalent Comparison Group Design

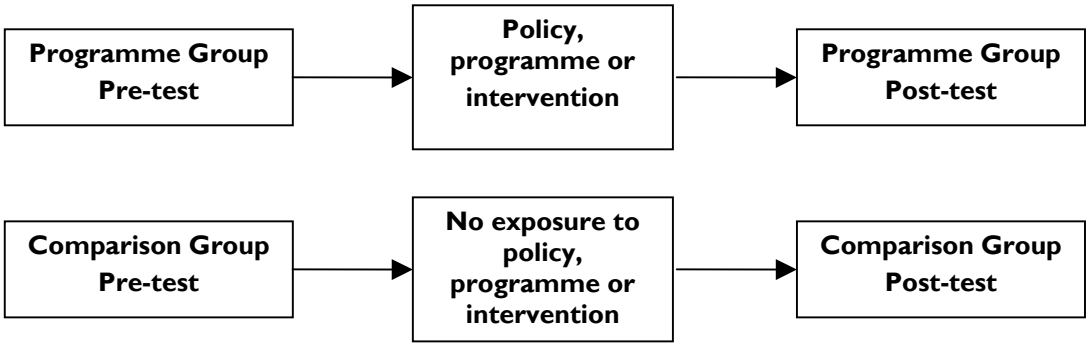


Figure 7.4 above illustrates a simple NECG design. Pre-test data and post-test data are collected for a group of units that receive or are exposed to the new policy or programme being evaluated. For simplicity, it is assumed here that pre-test data are pre-programme measures of outcome or dependent variables of interest. Post-test data are measures of the dependent variables or outcomes after the programme or policy has been introduced. Pre- and post-test data are collected for a comparison group and programme group at the same points in time.

The object of the comparison group is to help the evaluator interpret estimated programme impacts through providing a more convincing estimate of the counterfactual than that obtained through a single group design. The designation 'non-equivalent' means that the comparison group can be selected in any number of ways, with the exception that access to the programme cannot be determined through random allocation, which as we have seen ensures statistical equivalence between programme and control groups.

Such a comparison group can be selected to evaluate a policy or programme in the following circumstances (Purdon 2002):

- *Voluntary pilot programmes* – comparison samples can be constructed from individuals eligible for the programme but who choose not to take part.
- *Mandatory or saturation pilots* – here a programme is tested which is mandatory, in that all individuals eligible for the programme or policy in the pilot area are compelled to participate, or that the pilot is implemented 'full-scale', where no within area control or comparison group is possible. The latter design is often used where evaluators wish to measure scale effects. In such circumstances a comparison sample of similar individuals can be selected from geographical areas where the programme or pilot is *not* being tested. Alternatively, a comparison sample could be selected from a nationally representative sample survey should it posses

enough individuals with similar characteristics to those in the programme group.

- *Voluntary full-scale national programmes* – Where programmes or policies that are introduced universally are to be evaluated a comparison group can be selected from individuals that are eligible for the programme but who chose not to participate

7.7.1 Selection bias and NECG designs

Collecting pre-test data for both programme and comparison group allows the evaluator to examine whether the two research groups differ prior to the introduction of the policy or programme being tested in terms of pre-test values. Random assignment ensures statistical equivalence at baseline in terms of both measured and unmeasured factors. With an NECG design, no such assurance exists and thus it is important to explore the extent to which programme and comparison groups might differ. Such differences might not have been measured, and may indeed not be measurable. Where differences between units in the programme and comparison groups do exist, be they observed or unobserved, and these differences are statistically related to the outcomes of interest, a selection problem is said to exist. A selection problem can interact with other threats to internal validity – for example a *selection-history* threat, or a *selection-maturation* threat, or *selection-regression* to the mean.

Other than threats to internal validity - namely the selection problem - NECG designs suffer in the main from similar threats to statistical validity, construct validity and external validity, as is the case with social experiments.

Example 7.3 The Problem of Selection Bias

Selection bias relates to unobservables that may bias outcomes (for example, individual ability, pre-existing conditions). Randomized experiments solve the problem of selection bias by generating an experimental control group of people who would have participated in a program but who were randomly denied access to the program or treatment. The random assignment does not remove selection bias but instead balances the bias between the participant and non-participant samples. In quasi-experimental designs, statistical models (for example, matching, double differences, instrumental variables) approach this by modelling the selection processes to arrive at an unbiased estimate using non-experimental data. The general idea is to compare program participants and non-participants holding selection processes constant. The validity of this model depends on how well the model is specified. A good example is the wages of women. The data represent women who choose to work. If this decision were made randomly, we could ignore the fact that not all wages are observed and use ordinary regression to estimate a wage model. Yet the decision by women to work is not made randomly—women who would have low wages may be unlikely to choose to work because their personal reservation wage is greater than the wage offered by employers. Thus the sample of observed wages for women would be biased upward. This can be corrected for if there are some variables that strongly affect the chances for observation (the reservation wage) but not the outcome under study (the offer wage). Such a variable might be the number of children at home.

Source: Greene (1997), from Baker (2000).

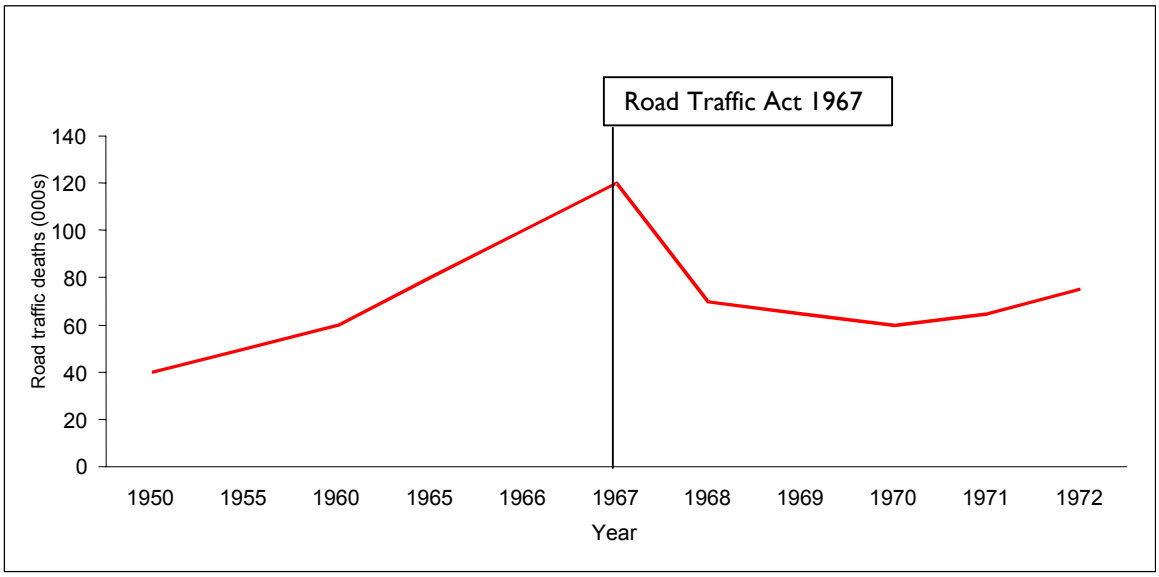
7.8 Interrupted time series designs

Interrupted time series designs investigate repeated observations of a constant variable over time and look for 'interruptions' to the series or sequence of observations (see Figure 7.5). Such interruptions might be attributable to an intervention, though they could also be a random blip in the series of observations (not likely in the example in Figure 7.5 where the interruption was continuous). There are different types of change to a time series sequence of observations; changes in the level and in the slope of the curve, the degree of permanency of the effect (as in Figure 7.5), and the type of impact (immediate or delayed).

In order to attribute the interruption in the time series of observations to a particular intervention it is important to know the specific point (e.g. the date) when the intervention was introduced. This must, of course, have been *before* the interruption in the observations if causality is to be inferred. If this is the case it is then necessary to consider, and rule out, any other *reasonable* explanations for why the interruption occurred. In the case of the clear interruption to the time series data on traffic accident fatalities in Figure 7.5 this would include ruling out, for instance, that there had been a fuel shortage after year 6; or that a major new tax had been introduced on road usage (or petrol); or that there had been a national alcoholic beverages strike; or that the price of alcoholic beverages had increased significantly, and so on. Assuming that none of these alternative explanations are accepted, we can infer that the noticeable reduction in road traffic accident fatalities was almost certainly attributable to the introduction of the Road Traffic Act.

It is also important when working with interrupted time series designs to establish that the variable(s) being measured are *constant over time*. Where definitions and counting practices change frequently over time (e.g. unemployment statistics) it is much more difficult, and sometimes impossible, to use such data as valid measurements, or to establish any causal significance to an interruption in the time series.

Figure 7.5 Interrupted time series design: road traffic deaths UK (1950 to 1972)



7.9 Statistical matching designs

Statistical matching designs are similar to the before and after non-equivalent control group method outlined above. Instead of finding a group of units whom we assume by virtue of some aggregate measure are a good match for the group receiving a treatment or intervention, we construct a comparison group through one-to-one matching. In other words, we attempt to find a control that matches a treated individual on the basis of what we observe about them. In short, matching attempts to "re-establish the conditions of an experiment [RCT/random assignment] when no such data are available" (Blundell and Costa Dias 2000: 444).

The basic idea with all forms of statistical matching is that the comparison group is so closely matched to the treatment group that the only difference between the two groups is the impact of the programme. Following on from this, the impact of the programme or intervention can be deduced from simple comparisons of means or proportions on the outcome variable (post-test) given samples of an adequate size. It is worth pointing out that in almost all cases impact estimates from statistical matching methods are likely to contain bias when compared to results from an RCT applied in the same context (LaLonde 1986; Lipsey and Wilson, 1993; Heckman, Ichimura and Todd 1997, Blundell and Costa Dias 2000). Some methods of matching however have proven to be better than others in replicating results from evaluations where random assignment has been adopted (Heckman, Ichimura and Todd 1997; Blundell and Costa Dias 2000). In general one-to-one statistical matching tends to perform better than non-equivalent control group

designs which are formed through natural assembly at the aggregate level[22].

Generally for statistical matching to be successful observations are required on a wide range of variables which are known from the research literature to be statistically related to both the likelihood that a unit will choose treatment and those related to the outcome measure.

7.9.1 Cell matching

Cell matching is slightly different to one-to-one matching. It involves dividing a sample of units who have received treatment into cells or strata. The cells are formed on the basis of responses to variables that are believed to influence the outcome variable of interest. For example, a sample of 1,800 treated units could be presented by cell as in Table 7.2.

Table 7.2 Example of a treatment sample broken down by cells for statistical matching purposes.

Area	**A**						**B**						**C**					
Sex	**Male**			**Female**			**Male**			**Female**			**Male**			**Female**		
Age group	0-16	16-65	65+	0-16	16-65	65+	0-16	16-65	65+	0-16	16-65	65+	0-16	16-65	65+	0-16	16-65	65+
N=	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

One would simply divide a sample of the eligible untreated population by the same strata to form cells as in Table 7.2. Selecting a particular cell, say all males aged between 0 and 16, living in area A, one would take the first treated unit in this cell and match them with a unit from the untreated sample in the same cell. This process of one-to-one matching would be repeated for every member of the treated sample, until one constructs a comparison sample of the same size[23]. The match can be made in the following ways within each cell:

- Select an untreated unit within the cell at random;
- Select an untreated unit who is closest to the treated unit (for example here, the untreated unit whose is nearest to the treated unit in terms of their age in months, or the untreated unit who resides close-by in terms of postcode); and

[22] In other words, one-to-one statistical matching should in theory perform better as a summative impact measure than control groups selected on an area basis, or naturally assembled unit such as a school class.

[23] Constructing a sample in this manner can proceed on the basis of either with or without replacement. Constructing a matched sample with replacement means that every untreated unit selected as a match for a treated unit is available for selection again as a match for a subsequent treated unit. Without replacement simply means that once an untreated unit has been matched with a treated unit, it not available for matching with any subsequent treated unit.

- Select an untreated unit by matching on a series of subsidiary variables within each cell.

One problem associated with cell matching, as with all forms of matching, is that one needs to know in advance which variables influence the outcome measure in order to control for their effect. This might require the researcher to undertake a review of the literature to determine the most important variables on which to match. There are also other potential problems with this approach. For instance, the matched data needs to be collected and this can often be expensive and time consuming if primary data collection is required. The process is also very data hungry in that you need a large amount of data in order to create sufficient matches.

7.9.2 Propensity score matching

Propensity Score Matching (PSM) can also be used to create comparison samples. The interested reader is referred to Purdon (2002) and Bryson, Dorsett and Purdon (2002) for an introductory discussion of propensity score matching within the context of evaluating labour market programmes. Briefly though PSM is a relatively new method it was first proposed by Rosenbaum and Rubin in 1983. It attempts, like all forms of matching, to match the treated and the untreated cases on a number of background characteristics so closely that the only difference between the two groups is the impact of the programme.

It is superior to cell matching because with the cell matching method it is often difficult or impossible to create matches when more than a few match variables are used. PSM solves this problem by matching, not on each and every individual characteristic, but on the overall effect of all characteristics. It does this by generating a predicted probability of participation in the treatment programme, using logistic regression. This probability is known as the propensity score. The propensity score represents the probability that a participant in the comparison group would have been selected for a treatment programme had the programme been running for this group.

In the logistic regression model the dependent variable equals 1 for programme participants and 0 for non-programme participants. A number of independent variables (covariates) can be used to predict programme participation, allowing more and better quality matches than would be the case with other forms of matching procedure.

As an example consider an offender treatment programme. Amongst cases in the treated group the propensity score represents the probability that an offender entered the treated group, and had they been selected for the programme, on the basis of the *covariates* alone. Thus if offender A in the treated group had a Propensity Score of 60% and offender B in the comparison group also had a PS of 60%, both offenders would have had an equal chance of being selected for treatment.

Having calculated the propensity scores for each offender in this way the evaluators would match each treated offender to an offender in the comparison group sample. There are several matching methods for doing this.

The most obvious method would be a 'one-to-one' match, whereby each participant must match a unique comparison group candidate on the propensity score. However, this matching criterion is too rigid because one may not be able to find an exact match. This is known as the 'support problem', whereby you can't find a supporting case in the comparison group to match the treated case. A possible solution to this problem would be to omit cases in the treated group for which one could not find a match (or common support) in the comparison group. However, this could result in a significant loss of cases and introduce selection bias. For example, some studies employing this method have lost over ¼ of treated cases.

As with all forms of evaluation method the level of attrition in the sample should be kept to a minimum. High levels of attrition, as noted elsewhere in this chapter, can undermine the external validity of an evaluation. For a researcher working in Government attrition undermines the policy relevance (i.e. the external validity) of findings because policy officials want to know the effect of the policy on all those who are eligible for the treatment, and not just a sub-sample of cases that could be matched. The ideal is to find common supports (or matches) for as many treated cases as possible. PSM attempts to deal with this problem using a number of different methods.

One solution used in PSM is to use the *nearest neighbour* matching method, whereby one takes each treated individual and matches them to a case in the comparison group with the closest propensity score. In addition, one can match a single case in the comparison group to someone in the treated group more than once. This avoids 'using-up' the comparison group cases, which would be the case if you permitted a comparison case to be used only once and thereby provides for more matches. The disadvantage of this approach is that it may produce some poor matches. For example, a comparison case may be the nearest neighbour to a treated case but they may have very different propensity scores. This problem usually occurs when one is dealing with small samples, which have a high degree of variation.

An alternative to the nearest neighbour method is caliper matching[24]. Caliper matching (Cochrane and Rubin, 1973) sets a tolerance range on the propensity score for a match (e.g., .01 to .00001). A non-participant is matched on the propensity score with a participant, provided that the difference in scores is within the tolerance range. This imposes a quality control on the match and so the treated and untreated groups will be similar. The problem here, however, is that it is not obvious *how* to set

[24] This is not the only alternative and the interested reader should refer to Purdon (2002) and Bryson, Dorsett and Purdon (2002) for further information.

the tolerance range. If the tolerance range is too narrow some cases will be lost while setting a broad tolerance range will reduce the quality of the matches.

7.9.3 Disadvantages of PSM

PSM is basically an improved version of cell matching, but with many of the same limitations. These can be summarised as follows:

- Large samples are required to create sufficient matches;
- The method can be very expensive because one needs to collect a great deal of information in order to create a predictive model of programme participation;
- Hidden bias may remain because matching only controls for observed variables.

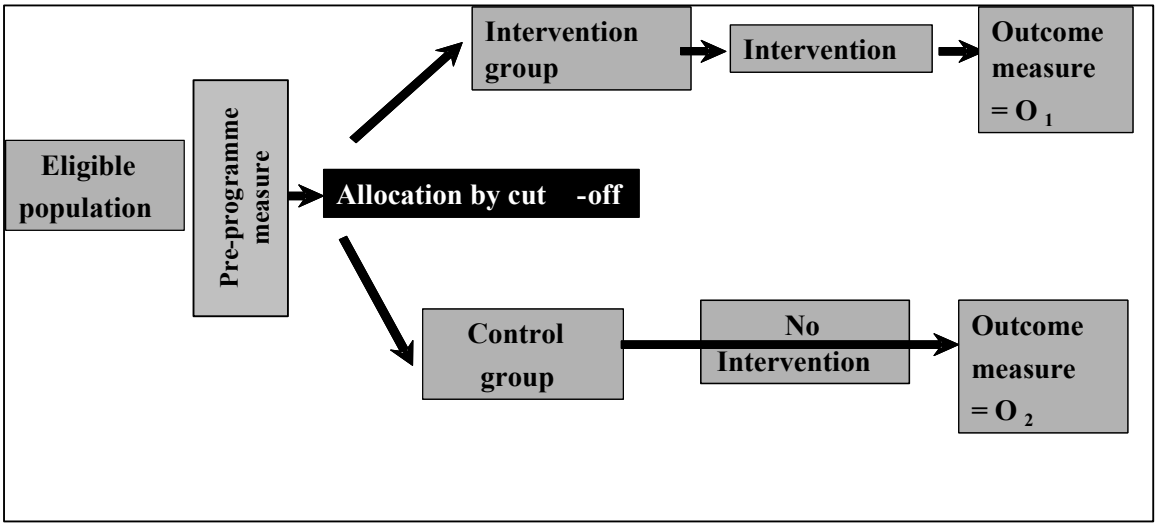
7.10 Regression discontinuity design

Regression Discontinuity Design (RD design) is a method that has been developed relatively recently but it is already established as the quasi-experiment that comes closest to an experimental design in eliminating selection bias (Mosteller, 1990). It has been described as "one of the strongest methodological alternatives to randomized experiments when one is interested in studying social programs" (Trochim, 1984).

The RD design is a type of *before-and-after two group design.* In other words all persons in the study are assessed before they receive treatment and are then re-assessed after receiving treatment. There are always two groups: an intervention group and a comparison group. In these respects the RD design is not unique. There are plenty of other designs, which could be called before-and-after-two-group designs. However, the unique feature of the RD design is the allocation process, whereby study participants are allocated to the control and treatment conditions *solely* on the basis of a cut-off score on a pre-programme measure. In other forms of quasi-experiment the allocation process is not controlled, and the treatment and comparison groups are self-selected. It is this feature that makes the RD design so much more robust than other forms of quasi-experiment.

In the RD design the *only* bias between the treatment and comparison groups is the difference in scores on the pre-programme measure. No other variable influences the selection process. This is not the case in other forms of quasi-experiment where only a limited number of variables are controlled and where an infinite number of unknown variables could influence the results of the evaluation. Moreover, in the RD design the source of the selection bias is not only known but it has been quantified by the pre-programme measure. This therefore allows for it to be controlled for by the use of regression analysis.

Figure 7.6 Regression Discontinuity Design



The use of regression analysis in the RD design can be best illustrated in Figure 7.7 and Figure 7.8

In Figure 7.7 participants are measured on the pre-programme measure (the assignment variable) and are then re-measured post test. In this hypothetical example those participants who scored above 49 points on the assignment variable received treatment while those who scored below this point did not. Each point on the graph shows the relationship between the pre and post measures for each person in the study. A regression line has been drawn between all the observations on the left hand side of the cut-off point (i.e. those who did not receive the treatment). This regression line is then projected through all the cases on the right-hand side of the cut-off point (the treatment side of the scatter-plot). It can be seen that the regression line predicts the post-test scores for both the observations on the left and right of the cut-off (i.e. for both the treated and untreated cases). This would imply that the treatment is ineffective because one would expect a *discontinuity* in the relationship between the assignment variable and the post-test score for the treated cases.

Figure 7.8 shows a regression line drawn through all the cases on the left hand side of the cut-off (i.e. those cases that did not receive treatment), and this line has then been projected through the treated cases. However, in this example the regression line does not predict the post test score of the treated cases. There is a *discontinuity* between the predicted and actual post-test scores for the treated cases – hence why the design is called regression discontinuity. This would indicate a treatment effect.

Figure 7.7 Regression Discontinuity Trial With No Treatment Effects

Discontinuity Trial With an Effective Treatment

Fig 7.1:Regression Discontinuity Trial With No Treatment Effects

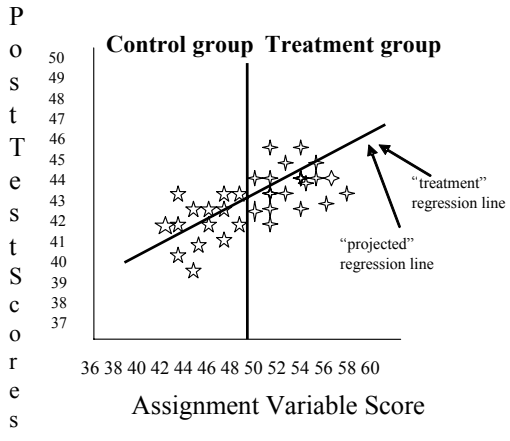
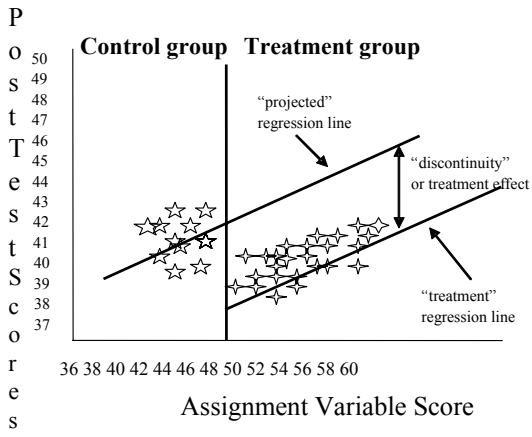
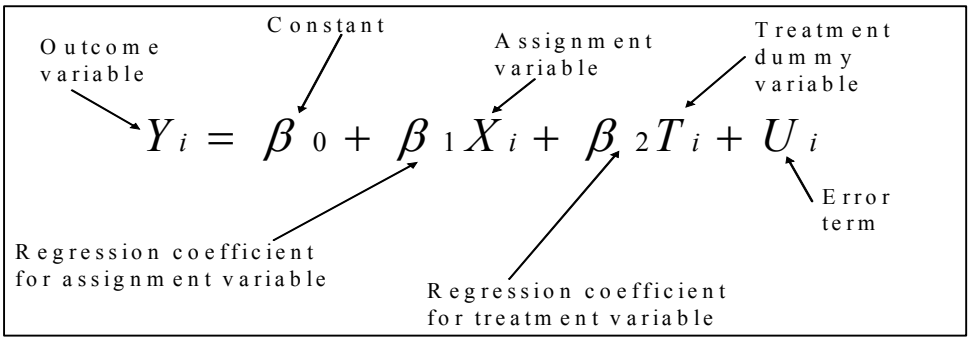


Fig 7.2: Regression Discontinuity Trial With an Effective Treatment



The regression lines in the above graphs represent a statistical model of treatment effect. The model can be specified by the following equation:

Figure 7.9 The specification for a regression model of treatment effect



Yi represents the outcome variable in the evaluation or, in other words, the post-test score. The model includes a constant value (B0) and a regression coefficient for the assignment variable (B1). The latter represents the relationship between the assignment variable and the post-test scores and therefore controls for the selection process (remember that the assignment variable was the only factor that determined whether or not a participant received treatment). The model then includes the regression coefficient for treatment (B2). This is computed by using a dummy variable which is scored 1 if the participant received treatment and 0 if they did not receive treatment. A value

significantly greater than zero for this coefficient would indicate a discontinuity in the slope of the regression line, representing an effect for treatment (which could be positive or negative).

7.10.1 Assumptions underpinning RD design

The regression discontinuity design can be characterised by three central assumptions (Trochim, 1984):

1. Perfect assignment. The assignment of participants to the comparison and the treatment group needs to be determined solely by the cut-off score on the assignment variable. No other factor should influence the selection process. A violation of this assumption, such as by allowing some participants to self-select group membership or by allowing a practitioner to allocate cases independently of the assignment variable, would introduce selection bias. This bias could not be controlled by the use of regression analysis because the bias would a) be unknown or only partially known b) unmeasured and so would not be amenable to statistical analysis.

2. Correct specification of the statistical model. It is important that the regression model accurately describes the true pre-test functional relationship. For example, if the relationship is linear but is modelled using logarithmic, quadratic, cubic or other regression functions biased estimates will result.

3. Absence of coincidental functional discontinuities. A factor that affects the post-test scores in one group but not the other could lead to a discontinuity that could be mistaken for a treatment effect. For example, if the treated group are located in one physical setting and the comparison group to another then the setting (and not the programme) may determine a change in post-test scores. The RD design therefore assumes that all group factors that differentially affect the post-test scores, other than the programme itself, are accounted for in either the design or the analysis (Trochim, 1984).

7.10.2 Data requirements

In order to implement a RD design the evaluator needs three essential pieces of information. These are as follows:

1. A pre-programme measure. This should be a quantifiable and continuous measure of some characteristic related to the post-test outcome that can be used to allocate participants to the comparison and treatment groups on the basis of a cut-off point on the scale. Each study

participant therefore needs to have a score on this measure.

2. Accurate data on who was allocated to the comparison and the treated groups (e.g. comparison group members=0 and treated cases=1).

3. A valid and reliable post-test measure designed to show the effect of the treatment programme.

7.10.3 Advantages of RD design

- The fact that persons are assigned to groups on the basis of a cut-off score on a pre-programme measure allows practitioners to target treatment at those most in need. For example, when there are a finite number of places on a treatment programme the allocation of places on the basis of need avoids some of the ethical issues associated with random allocation. The use of an objective scale to allocate treatment places is also transparent and so participants understand why they have been allocated or denied treatment.

- The main methodological advantage of the RD design is that the selection process is controlled, and the source of the selection bias (the assignment variable) is known and quantified. This is not the case with other forms of quasi-experimental design, which suffer from an unknown number of threats to their internal validity.

7.10.4 Disadvantages of RD design

- A RD design can only be applied in instances where it is possible to allocate participants on the basis of a pre-programme measure. Practitioners and/or participants may not want allocation to be determined in this way, and even if they did a valid and reliable pre-programme measure may not be available.

- The statistical analysis involved in RD design can be complex.

- Large sample sizes are required to generate sufficient statistical power to detect a treatment effect.

- The allocation process may be difficult to implement. For example, practitioners may not always allocate participants using only the cut-off score, or the assessment measure may be used inappropriately.

- The results of a RD design are more difficult to interpret and to communicate than in other forms of design.

It is fair to say that the opportunities to use this design are limited. The method though has been used to evaluate a number of policy programmes. For instance Berk and Rauma (1983) used a RD design to evaluate the effect of an offender treatment programme. In this instance the researchers estimated the effect of a programme that provided unemployment benefits to released prisoners in California.

The research team was able to implement an RD design because offenders were only eligible for payments if they had worked in prison for more than 652 days, and the amount of payment was proportional to the number of days worked. The formula applied was explicit and quantitative, the cut-off point was uniformly 652 days of work, and payments were made above and below that point (and only that point). The researchers then compared the re-arrest rates of those who received payments to those that did not, while using the hours worked in prison as a control variable (i.e. modelling selection exactly). The regression analysis showed that ex-prisoners who received payments were estimated to have 13 per cent fewer arrests. This estimate is unbiased insofar as the selection process is known and accurately represented in the statistical analysis by the variable 'hours worked in prison' (Rossi et al, 1999).

7.10.5 Philosophical issues

Social experiments (including quasi-experiments) are often criticised from an epistemological perspective, in that as a method, experimentation is commonly associated with 'positivism'. Positivism holds that the purpose of scientific enquiry is to test and predict the phenomena experienced in the social world, and that science should only be concerned with what can be measured. In a 'positivist' world, one cannot observe the process of cause but simply measure the consequences or outcomes of causal processes. According to the positivist perspective, the objective of the scientific process is to be able to predict and control the material world, and by extension in the social sciences, the social/political world.

Positivism, is seen by some critics, as a rather crude and naïve account of the scientific process and has been superseded most prominently by interpretivism, phenomenology, critical realism and post-modernism in the social sciences. A common argument of these critics is that science can never completely account for the nature of reality and that all scientific measurement is subject to various forms of error. Consequently all scientific theory, and all knowledge, is revisable and can be deconstructed and reconstituted. The idea that an evaluator can be 'objective' about the social world is rejected. Instead, individual evaluators must compare and contrast multiple accounts and attempt to *triangulate* their findings with those of others.

One of the more recent and influential critiques of random allocation comes from the 'realist' perspective outlined by Pawson and Tilley

(1997). They characterise the experimental method as being based on a 'successionist' conceptualisation of cause (similar to what Shadish, Cook and Campbell call *causal description*), whereby evaluators simply implement random allocation methods and neglect to study the processes or causal mechanism yielding the measured outcomes and impacts. Pawson and Tilley contrast the crude simplicity of experimentation with the subtle more nuanced conceptualisation of cause expounded in 'generative' theories of causation (similar to Shadish, Cook and Campbell's concept of *causal explanation*). In their view experimentation neglects the vital task of studying transformation directly and accounting for all the processes that bring about change. Pawson and Tilley are concerned with elucidating the context, mechanisms and outcomes (regularities) of policies and programmes. Part of their critique also highlights the practical difficulties associated with implementing social experiments and the claimed inconsistencies and variability of findings from them (Heckman and Smith, 1995, also make this point).

Although Pawson and Tilley (1997) and many others have extreme misgivings concerning the experimental method such approaches retain currency and usefulness. The reason for this is that the experimental approach, although inappropriate in some circumstances and never sufficient on its own, continues to provide policy makers, particularly those who have to make decisions regarding the allocation of resources, with the types of information they find useful. Few if any policy makers or practitioners would doubt that knowledge is contingent, ephemeral, and revisable in light of interpretation and further analysis. Nonetheless, most would be content to work with reasonably stable understandings of the social world in order to plan and predict within certain parameters of risk. Furthermore, there are examples internationally of where the experimental method of evaluation, within a multi-method approach, has been shown to provide cumulative, reliable findings that have been influential in the development of policy (Greenberg and Mandell 1990).

Social experiments are never, or should never, be used as the sole source of evidence in evaluating a programme or policy. Experiments need to be conducted alongside a thorough process study, which explicitly seeks to understand the context and causal processes being evaluated.

7.11 Summary and conclusions

This chapter has considered experimental and quasi-experimental methods for evaluating the effect of policies and programmes, and has detailed the advantages and disadvantages associated with these methods.

Social experiments essentially test whether a programme or policy has led to change in the outcomes that the programme or policy was

designed to affect, over and above that which would have occurred in the absence of the programme or policy. Social experiments use random allocation and so yield unbiased estimates of the programme or policy's impact. In other words they provide an estimate of impact that is entirely attributable to the programme or policy itself, rather than some other factor(s). In contrast, quasi-experiments are unable to eliminate the possibility that some extraneous variable may account for the impact of the programme. This is not to say that quasi-experimental methods should be ignored. Indeed there are circumstances when it is not possible to use random allocation for good practical and ethical reasons. In these instances a well implemented quasi-experiment, using a regression discontinuity design or a matched comparison design with PSM, provides a robust alternative.

The experimental method is not without its critics (see Pawson and Tilley, 1997). However, a structured and properly implemented design, when combined with a thorough process study, can overcome many of these criticisms and provide robust data on the effect of a policy or programme.

7.12 References

Baker, J. L. (2000) Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners. *Directions in Development. The World Bank. Washington, D.C.*

Berk, R. A. & Rauma, D (1983) Capitalizing on nonrandom assignment to treatments: a regression-discontinuity evaluation of a crime-control program, *Journal of the American Statistical Association*, March, 78(381): 21-27.

Bloom, H. S. (1995) Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs, *Evaluation Review*, 19: 547-556.

Blundell, R. & Costa Dias, M. (2000) Evaluation methods for non-experimental data, *Fiscal Studies*, 21(4).

Bonate, P. L. (2000) *Analysis of Pretest-Postest Designs*. Boca Raton, Chapman & Hall/CRC.

Boruch, R. F. (1997) *Randomized Experiments for Planning and Evaluation: A Practical Guide.* Thousand Oaks, Sage Publications.

Bryson, A., Dorsett, R. & Purdon, S. (2002) *The Use of Propensity Score Matching in the Evaluation of Active Labour Market Policies*, Department for Work and Pensions Working Paper No. 4 London, HMSO.

Burtless, G. (1995) The Case for Randomized Field Trials in Economic and Policy Research, *Journal of Economic Perspectives*, 9: 63-84.

Burtless, G. & Orr, L. L. (1985) Are Classical Experiments Needed for Manpower Policy?, *The Journal of Human Resources*, 21: 607-639.

Campbell, D. T. & Kenny, D. A. (1999) *A Primer on Regression Artifacts*. New York: The Guildford Press.

Campbell, D. T. & Stanley, J. C. (1963) *Experimental and Quasi-experimental Designs for Research*. Boston: Houghton Mifflin Company.

Clarke, M. & Oxman, A. D. (eds) (1999) *Cochrane Reviewer's Handbook 4.0*. Oxford, The Cochrane Collaboration.

Cochrane, W. & Rubin, D. (1973) Controlling bias in observational studies: *A Review, Sankhya*, Series A, 35: 417-46.

Cook, T. D. (2002) Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community has Offered for not Doing Them, *Educational Evaluation and Policy Analysis*, 24: 175-199.

Cook, T. D. & Campbell, D. T. (1979) *Quasi-experimental Design & Analysis Issues for Field Settings*. Boston: Houghton Mifflin Company.

Greenberg, D. H. & Mandell, M. B. (1990) *Research Utilization in Policymaking: A Tale of Two Series (of Social Experiments),* Discussion Paper No. 925-990. Institute for Research on Poverty, Madison.

Greene, W. H. (1997) *Econometric Analysis*. Hemel Hempstead, New Jersey, Prentice Hall Press.

Greenberg, D. H. & Shroder M. (1997) *The Digest of Social Experiments*. Washington: Urban Institute.

Heckman, J. J. (1979) Sample Selection Bias as a Specification Error, *Econometrica*, 47: 153-161.

Heckman, J. J. & Smith, J. A. (1995) Assessing the Case for Social Experiments, *Journal of Economic Perspectives*, 9: 85-110.

Heckman, J., Ichimura, H. & Todd, P. (1997) Matching as an econometric evaluation estimator: evidence from evaluating a job training programme, *Review of Economic Studies*, 64: 605-654.

LaLonde, R. (1986) Evaluating the Econometric Evaluations of Training Programs with Experimental Data, *American Economic Review*, 76(4): 604-620.

Morris, S., Greenberg, D., Riccio, J., Mittra, B., Green, H., Lissenburgh, S. & Blundell, R. (2003) *Designing a Demonstration Project – An Employment, Retention and Advancement Demonstration for Great Britain*. Occasional Paper No.1. London, Strategy Unit, Cabinet Office.

Oakes, M. J. & Feldman, H. A. (2001) Statistical Power for Nonequivalent Pretest-Postest *Designs, Evaluation Review*, 25: 3-28.

Orr, L. L. (1999) *Social Experiments: Evaluating Public Programs with Experimental Methods*. Thousand Oaks, Sage Publications.

Pawson, R. & Tilley, N. (1997) *Realistic Evaluation.* London, Sage Publications.

Purdon, S (2002) *Estimating the Impact of Labour Market Programmes*. Department for Work and Pensions Working Paper No. 3. London, HMSO.

Reichardt, C. S. (1979) The Statistical Analysis of Data from Nonequivalent Group Designs. In T. D. Cook & D. T. Campbell (eds) *Quasi-experimental Design & Analysis Issues for Field Settings*. Boston, Houghton Mifflin Company: 147-205.

Rosenbaum PR, Rubin DB. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70: 41–55.

Rossi, P. H., Freeman, H. E. & Lipsey, M. W. (1999) Evaluation: *A Systematic Approach*. Thousand Oaks, Sage Publications.

Scriven, M (1991) *Evaluation Thesaurus*: Forth Edition. Newbury Park, Sage Publications.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference.* Boston, Houghton Mifflin Company.

Smith, A. S., Youngs, R., Ashworth, K., McKay, S., Walker, R., Elias, P. & McKnight, A. (2000) *Understanding the Impact of Jobseeker's Allowance.* Department of Social Security Research Report No. 111. Leeds, Corporate Documents Services.

Trochim, W. M. (2000) *The Research Methods Knowledge Base*, 2nd Edition. Internet WWW page, at URL: <http://www.socialresearchmethods.net/kb/ (version current as of October 20, 2006).

Weiss, C. H. (1998) *Evaluation*. Upper Saddle River, Prentice-Hall.

Winship, C. & Morgan, S. L. (1999) The Estimation of Causal Effects from Observational Data, *Annual Review of Sociology*, 25: 659-707.

7.13 Appendix

Box 7.1 Planning Social Experiments – The Minimum Detectable Effect

The history of evaluating social programmes in North America and the United Kingdom suggests that the effects of social programmes often tend to be quite modest (Greenberg and Schroder 1997). Combined with the fact that individuals subject to social interventions tend to be relatively heterogeneous suggests that samples for measuring impacts have to be large in order to detect programme impacts. When setting out to design a social experiment, one of the most effective ways of ensuring that it will be able to detect a programme impact is to ensure that it has a plausible *minimum detectable effect* (Bloom 1995) and that the design is statistically 'powerful' enough.

$$MDE = z\sqrt{\frac{\sigma^2\left(1-R^2\right)}{p(1-p)n}}$$

The equation above provides a method for calculating the minimum detectable effect, or MDE, for a specified outcome, randomised design and sample design. The minimum detectable effect is expressed in the units in which the outcome is measured. So, for an outcome measured as a percentage point, it represents the minimum detectable percentage point effect that can be detected given the various assumptions represented by the arguments in the equation. Taking each of these in turn:

- 'σ^2' represents an estimate of the population variance of the outcome under consideration. Such estimates can be obtained from existing studies such as similar experimental or non-experimental studies, or from large probability surveys.
- '*p*' in the equation above represents the proportion of the sample allocated to the programme group, which is usually 0.5, where each eligible individual has an equal chance of being assigned to the programme group.
- '*n*' represents the sample size that it is anticipated will be available for estimating impacts. The calculation of '*n*' should take into account any potential non-response in the case of survey samples.
- '*z*' is the sum of z values drawn from a standard normal distribution for one minus statistical power (β) and one minus statistical significance (α). In other words the summation of $Z_\alpha + Z_\beta$ for a one tailed statistical test and $Z_{\alpha/2} + Z_\beta$ for a two-tailed statistical test. For further details see Box 7.2 below
- '$\left(1-R^2\right)$' acts to reduce the variance and thereby lower the MDE all things being equal where R^2 is non zero and represents the explanatory power of the regression equation used to estimate programme impacts.

To summarise, the minimum detectable effect describes the *sensitivity* of an experimental design in the units in which a given outcome is measured. The greater the values of '*z*' and 'σ', in other words the more heterogeneous the outcome in the population we are studying and the more precise and statistically powerful and precise we wish the results of our experiment to be, the greater the MDE (the less sensitive the design). The greater the value of '*n*' or the sample size and/or the larger the value of 'R^2', the greater the sensitivity of the design and the smaller the MDE. Experimental designs where the probability of allocation to programme and control group is equal are most sensitive all things being equal.

Box 7.2 Statistical Error Rates in Planning for Social Experiments

Experiments essentially attempt to falsify the null hypothesis that the programme or policy has no impact - that any observed differences between the two groups' results from sampling error. Statistical significance of 95 per cent, or a Type 1 error rate of five per cent, means that there is a five per cent chance of *failing to reject* a 'true' null hypothesis of no impact or effect (known as a false positive conclusion). In other words, where there is no true impact there is a five-percent chance of mistaking a difference between programme and control groups, resulting from sampling error, for a genuine impact. Statistical power of 90 per cent implies a 10 per cent chance of failing to reject the null hypothesis when it is in fact false and a true statistically significant impact exists (known as a false negative conclusion). **Statistical power is therefore defined as the probability of detecting a statistically significant impact where a true effect exists**.

Choosing the levels of Type 1 and Type 2 statistical error is important when designing a social experiment and thinking about an experiment's minimum detectable effects for various outcomes of importance (see Box 7.1 above). It is important to consider the costs that are likely to arise from making both types of error. For a discussion on this, readers are referred to Orr (1999). Likewise, in deciding whether hypothesis testing should proceed on the basis of a one or two-tailed statistical test, readers are referred to Bloom (1995).

A range of values for '*z*' – statistical power and significance

Statistical Power ('1-β', where 'β' is a Type 2 error rate)	**Significance level ('α' or Type 1 error rate)**		
One-sided hypothesis test	***0.10*** ($z_\alpha = 1.28$)	***0.05*** ($z_\alpha = 1.64$)	***0.01*** ($z_\alpha = 2.33$)
90% ($z_\beta = 1.28$)	2.56	2.93	3.61
80% ($z_\beta = 0.84$)	2.12	2.49	3.17
70% ($z_\beta = 0.52$)	1.80	2.17	2.85
Two-sided hypothesis test	***0.10*** ($z_{\alpha/2} = 1.64$)	***0.05*** ($z_{\alpha/2} = 1.96$)	***0.01*** ($z_{\alpha/2} = 2.58$)
90% ($z_\beta = 1.28$)	2.92	3.24	3.86
80% ($z_\beta = 0.84$)	2.48	2.80	3.42
70% ($z_\beta = 0.52$)	2.16	2.48	3.10

Notes: based on a table from Bloom H (1995: 550: Table 1)

GSR
GOVERNMENT SOCIAL RESEARCH
Analysis for Policy

The Magenta Book: guidance notes for policy evaluation and analysis

Background paper 8: how do you know why (and how) something works?
Qualitative methods of evaluation

Published: September 2004
Updated October 2007

Government Social Research Unit

HM Treasury
1 Horse Guards Road
London SW1A 2HQ

8 How do you know how (and why) something works?

8.1 Introduction – qualitative research: what is it and where does it come from?

Qualitative research encompasses a range of different approaches, disciplines and data collection methods. It includes research methods that capture naturally occurring data in their real-life context (principally participant and non-participant observation, documentary analysis, discourse and conversation analysis), and those that generate their own data through a reconstruction or retelling of views or behaviours (principally different forms of interviews, and focus groups).

Qualitative social research has its roots in the late nineteenth-century reaction against 'positivist' methods of social science and the attempts of the latter to emulate the principles and procedures of natural science. Rather than examine relationships in quantitative terms, or test hypotheses using statistical techniques, the interpretivist tradition argues that social research should provide in-depth, qualitative understanding of the subjective meanings of social life. In other words, it grew out of a concern to understand social life as it is experienced by people and as they make sense of it, rather than to derive laws or theories about it.

The natural sciences typically look for law-like patterns to social life – laws or theories which can be said to apply invariably. Interpretivist social science, on the other hand, looks to understand social phenomena from the viewpoint of the individuals and groups who experience these phenomena in specific (and not necessarily reproducible) social contexts. For interpretivists, social research is more akin to history than to natural science; i.e. it involves understanding, interpreting and explaining events, contexts and people in terms of personal and shared meanings, rather than looking for invariant laws detached from human agency and experience. Although qualitative research is usually associated with the interpretivist tradition and quantitative research with positivism, some commentators argue that this overplays distinctions in the ways in which they are used, particularly in multidisciplinary research, and that each can contain at least some elements of both positivism and interpretivism.

8.2 Key features of qualitative research

Qualitative social research is diverse. As a research approach, it draws on a number of different philosophical schools of thought and academic disciplines, particularly sociology, anthropology, philosophy, linguistics and psychology. Qualitative research has also developed a range of research methods and techniques (for example, in-depth interviews and focus groups). Qualitative methods are usually employed in *naturalistic*

settings, rather than the contrived environments of controlled trials or social surveys. The goal of most qualitative social research is to capture as closely as possible the understandings, interpretations and experiences of ordinary people in their everyday lives and environments.

Qualitative social research is also used with quantitative methods, and is undertaken in less than wholly naturalistic settings (e.g. as part of a social survey or a controlled evaluation). One of the major challenges of qualitative social research is to capture and explain social life as authentically as possible without affecting, manipulating or changing it. This is often very difficult. Qualitative social research, like most other types of social research, brings to the topic that it is investigating theories, background knowledge and the researchers' own experiences and meanings, all of which influence, if not bias, the topic or subject of inquiry. This raises important methodological problems about the status of the knowledge and data gathered by qualitative social research. Questions about the validity, reliability and generalisability of qualitative data are just as important as they are for quantitative and experimental research, though these issues are generally discussed by qualitative researchers using a different vocabulary (e.g. credibility, defensibility, external validity – see Spencer *et al*., 2004).

Given these different origins and approaches, it is unsurprising that there are few shared definitions of qualitative research. There is, however, broad agreement about its key characteristics, although the emphasis given to individual features will vary (Spencer *et al*., 2004, drawing on Bryman, 2001; Denzin and Lincoln, 2000; Hammersley and Atkinson, 1995; Mason, 2002; Patton, 2002). Generally, the key characteristics are seen as being:

- in terms of approaches
    - a concern with *meanings*, especially the subjective meanings of participants;
    - a commitment to viewing (and sometimes explaining) phenomena *from the perspective of those being studied*;
    - an awareness and consideration of the *researcher's role and perspective;*
    - *naturalistic inquiry* in the *'real world'* rather than in experimental or manipulated settings;
    - a concern with *micro*-social processes (i.e. their manifestation at the level of individuals, groups or organisations);
    - a mainly *inductive* rather than deductive analytical process (i.e. broadly, deriving theories or findings from empirical research data, rather than deducing a hypothesis a priori which is then tested by empirical research).

- in terms of research conduct
    - prolonged *immersion* in, or *contact* with, the research setting;
    - the absence of methodological orthodoxy and the use of a *flexible (emergent) research strategy*;
    - the use of *non-standardised*, *semi-structured or unstructured methods of data collection* which are sensitive to the social *context* of the study;
    - the capture of data which are *detailed, rich and complex*;
    - the collection and analysis of data that are mainly (although not exclusively) in the form of *words and images* rather than numbers.
- in terms of the use and presentation of research data
    - the setting of data in *context;*
    - a commitment to retaining *diversity and complexity* in the analysis;
    - a respect for the *uniqueness* of each case as well as themes and patterns across cases;
    - attention paid to *categories* and *theories which emerge from data* rather than sole reliance on a *priori* concepts and ideas;
    - *explanations offered at the level of meaning* (i.e. the individual and shared meanings that things hold for people) or in terms of local 'causality' (why certain interactions do or do not take place in individual cases) rather than context-free laws of general application.

8.3 Applications of qualitative research to evaluation

These key features of qualitative research mean that it makes a distinctive contribution to policy evaluations, particularly because of its ability to explore issues in depth and capture diversity, its concern with context, and its focus on exploring meanings. This means that it can bring real depth to the understanding of the contexts in which policies operate and their implementation, processes and outcomes.

Personal, social, structural and environmental contexts can all be important in policy evaluation. To illustrate, in an evaluation of a welfare to work policy aimed at disabled people, qualitative research would provide understanding of how disabled people understand and experience disability, the values associated with work and social security benefits, and feelings about and experiences of working. These individual contexts would provide critical understanding of how people relate to

or think about the policy or service. Qualitative research would also illuminate issues such as how employers' attitudes or policies might act as barriers to disabled people who want to work, how the social security system can complicate or hinder the move from benefits to work, or how the culture of an organisation delivering the service shapes its approaches to implementing it. Understanding these broader aspects of the setting in which the policy works is also critical. The ability of qualitative research to illuminate contexts means it plays an important role in the development of policy. By understanding how people view an aspect of their world, and how they behave or make decisions within it, qualitative research can help to identify possible policy approaches and strategies, and to enhance the likely success of their implementation and delivery.

In terms of implementation, qualitative research can be used to look at issues such as what and who is involved in the process of implementation, the steps and processes involved, barriers and facilitators, decisions made, whether the policy is implemented as envisaged, and the reasons for deviation from the original design.

Qualitative research also lends itself to evaluations which require an understanding of processes. It can, for example, generate a detailed description of what interventions are involved in a service or policy, who provides them, what form they take, how they are delivered, and how they are experienced by participants and by those who deliver them. It can provide an in-depth understanding of the decisions, choices and judgments involved, how they are made and what shapes this. This is particularly important where the policy or intervention is itself highly process-orientated, where the intention is to effect change through interactions (for example, between several members of staff and a client) rather than through a one-off event or input.

Qualitative research also plays a key (although sometimes neglected) role in understanding impacts and outcomes. Rather than providing quantitative measurements of gross or net impact, it can answer more detailed questions which might be summarised as *'how, under what circumstances, in what ways and for which types of people is the policy working … and what do we mean by "working" anyway?'* It can tell us about the range and types of impacts a policy has, giving a voice to outcomes that were not anticipated or intended and which an evaluator might not have thought to consider. In this respect, qualitative research and evaluation has much in common with the theories of change and realist approaches to policy analysis.

Qualitative research can illuminate the meaning that outcomes hold, or how they are perceived and experienced, by the people involved. Building up from detailed individual accounts, it tells us how outcomes are arrived at – which elements of the programme, at which stages, experienced in what ways, and with what relationship to changes in a client's wider life, contribute to the impacts experienced. It explains

what might act as facilitators or barriers to the desired impacts, how they act in this way, and how barriers can be overcome and facilitators harnessed. And it explains how and why the impacts vary.

Example 8.1 Fathers in Sure Start Local Programmes (Lloyd *et al.,* 2003)

> The Sure Start programme provides a good example of how qualitative research is used alongside quantitative evaluation to provide a rounded understanding of how this programme is working. In addition to a national survey of 250 Sure Start areas and 50 Sure-Start-to-be areas in order to assess the impact of the Sure Start programme, the evaluation team is also undertaking a series of themed studies and local context analyses. By using in-depth interviews, informal observations and documentary sources, qualitative research identified, for example, that fathers were a hard-to-read group for Sure Start and other community-based programmes and that their needs were not always recognised or accommodated by these government programmes. It was also able to identify successful strategies for engaging fathers in Sure Start programmes.
>
> http://www.surestart.gov.uk/

Qualitative research is particularly valuable in certain evaluations, for example (Davies, 2000; Patton, 2002; Popay and Williams, 1998; Williams, 1986):

- where a policy or social context is not well understood, and the evaluation questions, issues or criteria are not immediately obvious;
- where 'insider' values and perspectives are particularly important as well as the 'official' perspective;
- where diversity in how the policy operates across different sites or services needs to be understood;

It can

- provide new insights into the implementation or experience of the policy;
- check for unintended or perverse consequences of the policy or project;
- explore the complexity of what goes on, in its natural settings;
- explore 'taken for granted' practices, lay behaviour and organisational cultures.

8.4 Implications for government research, evaluation and policy making

Qualitative research will make an important contribution to any policy evaluation where what is required is more than a quantification of users, costs, outcomes or estimated impact. Given that qualitative research is particularly helpful in understanding *why, how,* and *under what conditions* policies, programmes and projects work or fail to work, it helps policy making by identifying:

- the context in which the policy operates, and what this means for the design, development and likely success of the policy;
- how the policy is delivered and experienced on the ground;
- what impacts it has and which aspects of the policy contribute to them;
- and it will provide a critical explanation of why the policy works for some people, or in some circumstances, but not for, or in, others.

It is invaluable in framing policy questions in ways that are meaningful to ordinary people, and in eliciting their perceptions, understandings and experiences of policy interventions. It is probably not too grandiose to suggest that qualitative research helps enhance the democratic process of government by introducing the citizens' (or 'users") perspective in rigorous, systematic and non-anecdotal ways.

8.5 Qualitative research methods

This section discusses the main qualitative research methods used in evaluation: in-depth interviews; focus groups and consultative and deliberative methods; participant observation and ethnography; documentary analysis; narrative and biographical approaches, discourse and conversation analysis and case studies.

8.6 In-depth interviews

8.6.1 What are they?

In-depth interviews (also called unstructured interviews) are probably the most frequently used form of qualitative research in government evaluation (Spencer et al., 2004). Personal spoken accounts are seen as having central importance in social research because of their power to illuminate meaning (Hammersley and Atkinson, 1995). Individual, personal accounts display the language that people use, the emphases they give, and allow people to give explicit explanations for their actions and decisions. The in-depth interview is sometimes described as being akin to a conversation, or a *'conversation with a purpose'* (Webb and Webb, 1932: 130). However, although a good in-depth interview will

appear fluent, spontaneous and naturalistic, it is in fact a much more one-sided and carefully managed event than an everyday conversation.

In-depth interviews are a fairly lengthy, focused and usually private type of interaction. As such, they provide an opportunity to collect rich and detailed data, with the research interviewer 'mining' the subject and encouraging the participant to give more and more depth to their responses. They are ideal for an in-depth exploration of a subject which gives the researcher a detailed insight into the participant's own world, their beliefs, experiences and feelings, and the explanations they provide for their own actions or beliefs. Such interviews also lend themselves well to exploring complex processes or unpacking decision making. A good research interviewer will build a rapport with the participant which particularly aids the exploration of sensitive, painful or difficult subjects. Biographical research approaches use particular forms of in-depth interviews and are discussed in section 8.12.

The degree of structure in interviews varies. A key feature of qualitative interviews is that questions are not pre-formulated: there is scope for the participant to influence the direction and coverage of the interview. In some studies, perhaps particularly where the purpose is to reveal to the researcher a social world that is unfamiliar to them, the interview may be relatively unstructured, with the researcher asking very broad questions and the participant shaping the account. In others, the researcher will have a stronger sense of the issues that need to be explored and will play a more active role in directing coverage. The terms 'unstructured' and 'semi-structured' are sometimes used to denote different degrees to which the research directs the interview, although there is not always consistency in the ways they are applied.

8.6.2 How are they conducted?

The process of generating data through in-depth interviews is both systematic and flexible. It is systematic in that there needs to be careful and detailed thought initially about the type of data required and how to generate or collect it, and some consistency between interviews in the issues covered. However, data collection also needs to be flexible to reflect the uniqueness of each individual case, to explore what is of particular relevance to it, and to allow the formulation of the research questions to develop and sharpen as the study proceeds.

Some form of instrument – usually described as a topic guide, interview guide or interview schedule – is required. These act as an *aide mémoire* in the field, and help to ensure that there are no gaps in interview coverage. They are also an important aspect of accountability to those funding, commissioning or steering research, and provide a public document of an aspect of the research process that it is otherwise difficult to describe. Research reports usually show the topic guide used – for examples, see some of the studies quoted in this chapter. However, topic guides should be seen as the starting point for data

collection, not as a straitjacket. They need to be used flexibly if the aim is in-depth exploration, and it is good practice to review and make any necessary amendments to the guide after the first few interviews.

Good interviewing involves asking a range of different types of questions. Wide, 'mapping' questions are generally used to open a research topic, but the participants' initial responses are followed up with further questions and probes to 'mine' the topic in depth. A variety of follow-up questions will be needed for each topic to amplify, clarify, and seek explanation, and to stimulate, challenge or encourage the participant to review the topic from other perspectives. For example, a 'mapping' question might be along the lines of 'How did you come to decide to use the service?' The respondent might reply that they thought it sounded as if it might be helpful. 'Mining' questions would then be asked to find out in what way the respondent thought it might be helpful, what they knew about it, and what sort of help they wanted from it, with probes such as 'why was that?', 'in what way?', 'how did you know that?' and so on used to ensure each issue is explored in depth. Although broad and open questions are critical, good in-depth interviewing also involves a high level of specificity to get an in-depth understanding of the participants' attitudes, experiences or behaviour. What is always important is that questions are clear and not leading: questions are formulated as neutrally as possible so as not to imply that a particular response is expected or required.

In-depth interviews are usually tape-recorded. Note-taking is not sufficient to capture data in enough depth: the nuances and details which constitute the richness of interview data are lost, and the effort required in keeping pace with the respondent impedes the interview interaction. There is a potential disadvantage that being tape-recorded changes what a participant says, or makes the interaction more formal and charged. In practice, people appear quickly to get used to the presence of the tape-recorder – they seem surprised, for example, when the machine clicks off after one side of tape. Using field notes to record what is heard or seen outside the main research interaction can also be useful.

8.6.3 Implications for government research, evaluation and policy making

In-depth interviews are a very rich resource for researchers and policy makers. They provide detailed personal accounts of beliefs or experiences. They display the language, constructions and emphases of people in ways that are very revealing. And they allow participants to give explicit explanations for their views, decisions or actions, describing what has shaped them.

There are some things that interviews cannot do. The emphasis is on the participant's own interpretation, so they may not be the optimal research method where people are unable or unwilling to give an open account of themselves, or where they find introspection and self-

questioning particularly difficult. Where it is difficult for people to reconstruct and retell something or where they are less likely to do so honestly or accurately, other methods of data collection, such as observation, would be more appropriate. There will also be limitations on the extent to which people are able to pinpoint what has led to a particular behaviour, or what difference an intervention has made, or what would have happened in its absence. Asking people about their perceptions of these things will always be important and illuminating, but their perceptions may not always fully reflect what would have happened if they had, or had not, experienced the policy. Surveys collecting quantifiable data and using experimental designs such as control groups or randomised control trials are therefore needed to give an accurate measurement of net impacts, although qualitative research works effectively alongside such methods to provide explanations for the impacts found.

But interviews have many uses in evaluations. For example, they can be used to find out:

- how a service was set up and delivered by the management team;
- how staff work, the decisions and judgments they make, what they think works well and less well;
- how clients experience a service;
- within the detailed context of an individual case, what impact a service has, and how.

It can be difficult to assess whether the interview process is being well conducted. Ultimately, the proof is in the pudding – whether the data generated is rich, in-depth and illuminating. But research managers and policy makers should expect to see:

- a topic guide or interview schedule which comprehensively covers the ground but clearly allows for flexibility;
- reflective preparation for fieldwork by the team;
- tape-recording and verbatim transcriptions of interviews;
- an account, if requested, of how the interviews were conducted, what emerged spontaneously, how issues were raised and pursued by the research team;
- evidence of depth and insight in the data collected, demonstrated in the way it is analysed and presented.

8.7 Focus groups

8.7.1 What are they?

Focus groups or group discussions usually consist of around four to eight people, who may be acquainted with each other or may be strangers, brought together to discuss a particular topic or set of topics. The discussions typically last around an hour and a half, although this is certainly not fixed (see below). The group is moderated or facilitated by a researcher. Although focus groups have acquired a somewhat dubious image, they are a well-established and rigorous method of social research and evaluation.

In focus groups, data are shaped and refined through the group interaction. Hearing from other participants stimulates further thought, encouraging people to reflect on their own views or behaviour and triggering further material. Focus groups are synergistic (Stewart and Shamdasi, 1990) in the sense that the group works together, and the group forum is used explicitly to generate data and insights (Morgan, 1997). They also provide a strong social context to the discussion. This may be a natural social context if those in the group already know each other (for example, colleagues). But even if the group members are strangers brought together for the research, there will be more spontaneity than in an individual interview. People's social frames of reference will be more on display, there will be insights into how ideas and language are shaped by the social context, and social constructions – normative influences, collective as well as individual self-identities and shared meanings – will be illuminated (Bloor *et al*, 2001; Finch and Lewis, 2003; Krueger and Casey, 2000).

Focus groups have an application in any study where what is sought is refined and reflective discussion, or the social context made visible. The data they generate is in depth, not at the individual level as with interviews, but because it is the result of listening, and thinking further. They provide opportunities for creative thinking, for projective or enabling techniques, for group work and for giving information, for example, on technical subjects. They can work very well in tackling abstract or conceptual topics, whereas on a one-to-one basis a participant may 'dry up'. They can also be used for sensitive subjects, provided there is enough similarity between participants in their social characteristics and their connection with the research subject to create an environment that feels safe.

Focus groups work well in combination with interviews or other research methods. For example, at the beginning of a study they can be used to map out the territory, to give early insight into how people approach, discuss and construct a subject. At the end of a study, they offer a deliberative forum for refining understanding of an underlying theme, exploring causes or origins of problems, examining implications

for service delivery or policy development, or generating or prioritising solutions.

Example 8.2 Attitudes and Aspirations of Older People (Hayden *et al.*, 1999)

This study formed part of a programme of research on older people by the Department of Work and Pensions. The qualitative study was preceded by a literature review on attitudes and aspirations of older people.

This research used 15 focus groups of between six and eight older people aged 50 and above across the UK. 20 follow-up in-depth interviews were conducted to develop case study data. The aim was to explore the views older people have about their lives and their expectations and wishes for the future and to improve the understanding of the factors influencing the attitudes and aspirations of older people.

The focus groups were designed to cover diversity in gender, ethnicity, dependency and location. The focus groups explored themes around employment, work and volunteering; around retirement and activities in retirement; on issues of health and social care; and on services, consultation and citizenship. The research confirmed that attitudes and views of older people are influenced by a variety of socio-economic factors and different life experiences. It found considerable agreement among all participants in their desire to be as active and independent as possible and identified a number of barriers to active ageing.
http://www.dwp.gov.uk/asd/asd5/102summ.asp

8.7.2 How are they conducted?

The role of the researcher or facilitator is key to making focus groups effective (Finch and Lewis, 2003). As with in-depth interviews, the amount of structure will vary. The researchers will, at least once the group is underway, want as much as possible of the discussion to emerge from the group itself. But they will also help the group to focus and structure their discussion, bring discussion back or move it on, widen the discussion to include everyone, and ensure a balance between participants. They guide and pace the discussion to ensure all the issues are covered, and they probe individuals and the group as a whole to encourage in-depth exploration. They are also alert to non-verbal language and to the dynamic of the discussion, and they need to challenge or stimulate the group if what is said seems too readily to reflect social norms or apparent consensus.

Careful consideration needs to be paid to the composition of a focus group. This includes how many focus groups need to be convened to cover an issue adequately (see Sections 8.5 and 8.6), and which combination of individuals in each focus group will work best. For example, it may be difficult for individuals in an organisation to express their views openly in a focus group setting if more senior members of staff from the organisation are also in the group. Similarly, it can be

difficult to discuss certain topics in mixed gender groups or mixed age groups.

Practical arrangements are also important – the location and timing of the group's meeting can make an enormous difference to who is willing to attend, and the research team needs to take steps to make participation easy for people and to encourage and reassure them about attending.

8.7.3 Implications for government research, evaluation and policy making

Focus groups are very useful for any evaluation question where deliberation, discussion and the stimulation of other people are likely to bring depth to the topic. They are not appropriate where the presence of others is likely to inhibit or constrain people, or influence them in ways that make their accounts less open or less reliable. They provide much less depth at the individual level, and so are less appropriate where the personal context is critical, or for very detailed or complex subject matter. As with in-depth interviews, the emphasis is on the person's own interpretation, which may not always be what is required. But they are ideal where hearing what others say helps people to sharpen their understanding and articulation of their own views or experiences – for topics that are a little abstract or not at the front of people's consciousness, for example, for illuminating different views or beliefs, or for highlighting variation in behaviours or decision making.

As with interviews, it can be difficult to assess the conduct of fieldwork, but the research manager or policy maker can expect the research team to be able to describe the approach they have taken, or plan to take, to carrying out the groups. They should be alert to signs in the presentation of the data that the group dynamic inhibited open discussion and should expect to see diversity, depth – and the unexpected.

8.8 Consultative and deliberative methods

8.8.1 What are they?

Since the latter part of the twentieth century there has been a considerable expansion in the range of methods used for consultative purposes, particularly within local government. The boundaries between consultation and qualitative research are not absolutely clear cut, and some consultation methods involve the application of established research methods in a more dialogic and deliberative process. Lowndes *et al.* (1998) found an increasing repertoire of methods being used in local government, with a growing interest in innovative methods as well as, or in place of, traditional ones. They list 19 forms of public participation used, including meetings, committees and forums;

interactive websites; surveys, polls and referenda; and citizens' panels and citizens' juries. More innovative forms of consultation are:

- the Delphi method (Adler and Ziglio, 1996; Cantrill et al., 1996; Critcher and Gladstone, 1998). This is an iterative process particularly aimed at forecasting, in which a group of experts are asked to respond individually to questions (either in a survey or using qualitative research). Initial responses are then circulated among panel members who are asked to review their own responses, with further rounds of circulating and refining responses until consensus, or agreed difference, is reached. The group does not meet physically.

- the Nominal Group Technique. A variant of the Delphi method, this follows a similar pattern to the Delphi method in terms of eliciting initial responses from panel members. After the first round of elicitations, however, further iterations are conducted using an interactive group format somewhat similar to a focus group. The aim of the group is to reach a consensus in areas of both agreement and disagreement. A good facilitator is vital to ensure that all participants in the group are able to contribute to the discussion, and that the group works in a disciplined manner to reach agreement and agreed difference.

- citizens' juries (Coote and Lenaghan, 1997; Davies et al., 1998; White *et al*., 1999). A group of between 12 and 20 people are brought together over the course of several days to hear from and put questions to expert 'witnesses', deliberate and discuss among themselves, and make recommendations about courses of action, which may or may not be consensual.

- deliberative polls (Fishkin, 1995; Park *et al.,* 1999). This is a set of activities with a focus on exploring how public views change when the public has an opportunity to become well informed on an issue. A survey is conducted to 'benchmark' public opinion. Participants attend a joint event, usually over a weekend, which involves discussion in small groups, expert plenary sessions and political sessions in which party spokespeople respond to questions. The survey is repeated to measure the direction and level of change in views.

- consensus conferences or workshops (Seargeant and Steele, 1998). These follow a model developed in Denmark in which a panel of around 16 people work together to define the questions they wish to address within a particular subject, question experts, receive information, deliberate, and aim to reach consensus. The panel produces its own report, which is presented to an open conference and discussed further.

- 'participatory appraisal'. This has historically been used in overseas development work but is now more common in the UK. It is designed to involve people, particularly from socially excluded communities, in decisions that affect their lives. It combines a variety of visual (mapping) tools with group discussions and semi-structured interviews (Oxfam UK, 2003).
- 'planning for real'. This is a community consultation process where models are used to encourage residents to explore and prioritise options for action (Gibson, 1998).

8.8.2 Implications for government research, evaluation and policy making

There is a wide range of consultation methods available to the government researcher and policy maker. This raises the question of when different methods should be used, and how the choice between different methods should be made.

Consultation methods have the potential to enhance policy formulation, evaluation and development, as well as to contribute to cultural and personal change within the organisations using them. Researchers and policy makers will want to look to consultation methods, rather than research methods, when they want to move beyond exploring people's views and behaviours, to getting them to come up with, or to appraise, solutions and strategies. The group orientation of most consultation methods means that they are particularly useful when the issues involved are technical, complex and require consensus, and where additional information needs to be fed into the process of deliberation.

Effective consultation requires:

- clarity of purpose;
- clarity about who should be consulted and what they are expected to contribute;
- organisational capacity and skills;
- careful selection of a consultation method which fits the question and is feasible given time and resources;
- careful planning and management;
- the desire and competence to make use of the outputs;
- evaluation of the method used (Audit Commission, 2003; Lowndes *et al*., 1998 Seargeant and Steele, 1998).

Consultation generally involves intensive exercises with relatively small groups, and thus raises questions about value for money and representativeness. Well-conducted consultation will help to highlight and explain areas of difference, as well as agreement, among participants.

A careful balance needs to be struck between the need for consultation to point to an agreed way forward and the danger that it produces an artificial consensus.

8.9 Participant observation and ethnography

8.9.1 What are they?

One of the major ways in which social research can understand an activity, group or process is to get as close as possible to them without disturbing its 'natural' operations (Hammersley and Atkinson, 1995). This can be done at one extreme by being a wholly detached observer of a social situation, working as unobtrusively as possible, making observations, listening, and remembering details. At the other extreme, one can join the group or activity in question and participate in it as a member in order to learn about it from the inside out. This may or may not involve 'going native', i.e. becoming so closely involved in the group, activity or processes that one loses one's detachment and outsider status. Clearly, there are positions mid-way between these extremes; one can function as an observer-participant or as a participant-observer, the difference being the degree of detachment and involvement that is possible for the social researcher (Figure 8.1)

Figure 8.1 The observer–participant spectrum

Social research has a fine tradition of using participant-observation/observer-participation to better understand gangs, criminal behaviour, drug use/misuse, school participation and achievement, health and illness behaviour and many other substantive topics. Two of the 'classics' of participant observation are W.F. Whyte's (1955) *Street Corner Society* and Elliot Liebow's (1967) *Tally's Corner*. Erving Goffman's work, *Asylums* (1961), *Stigma* (1968), *Presentation of Self in Everyday Life* (1959) and *Behaviour and Public Places* (1963) stands in the pantheon of participant-observational research.

8.9.2 Ethnography

Ethnography is a method used by anthropologists which has been adopted by social researchers more generally. The term 'ethnography' means the description ('*graphy*') of a people or culture ('*ethno*'). More precisely, then, ethnography is the *detailed description of a culture, group or society, and of the social rules, mores and patterns around which that culture,*

group or society are based. Ethnographic inquiry is able to elicit the *cultural knowledge* of a group or society; i.e. what one needs to know to be a fully competent member of that society. Ethnography also involves the detailed investigation of the *patterns of interaction* of a group or society in order to understand the values, processes and structures (including language) of that group.

As with other types of qualitative social research, ethnography studies social groups and social activity in as 'natural' a way as is possible, i.e. without trying to affect or manipulate them in any way. Ethnography also shares with other types of social research an interest in capturing the 'first-order' meanings of the people it is studying, and their relationship to the 'second-order' categories and constructs that are used by researchers. Ethnography is a social research method that allows researchers to 'get inside' a group, culture or society and understand it from the inside out (Chambers, 2000).

8.9.3 How are they conducted?

Observation, listening, remembering and detailed note-taking are key techniques for social researchers using participant-observation and ethnographic methods of inquiry. A good participant-observer or ethnographer will learn to observe what is going on in a group, or a social situation, and to identify verbal and non-verbal communication, patterns of interaction, people's responses to verbal and non-verbal activity, body language, people's demeanour and deference to others, status hierarchies, and the like.

There is some debate among participant-observers and ethnographers as to whether one should take notes and make recordings of what is observed *in situ*, or whether one should refrain from doing so until the activity being observed has finished. If notes are not taken during observation, the observer/participant will make detailed field notes and recordings of what they have observed as soon as possible after the observation has been concluded, usually in the privacy of a private room or study.

In the past two decades or so, audio- and video-recording has become more readily available to social researchers, and is becoming less intrusive and more discrete. This has allowed participant-observers and ethnographers to collect audio- and video-taped data on everyday social settings such as classrooms, doctors' surgeries, courtrooms, and office life. Audio- and video-taped data has the advantage of allowing extensive (some say exhaustive) analysis of naturally occurring social activity, and for these activities to be retrieved for further analysis, verification and challenge by others. This can enhance the reliability and validity of qualitative analysis, and provide more transparency of qualitative social research.

Against these advantages is the potential disadvantage that audio- and video-recording disturbs the 'natural' social activity one is trying to

capture and understand. To some extent this can be minimised over time as people's awareness of the presence of audio microphones or video cameras usually diminishes as they get used to the presence of these devices. In the short term, however, and possibly beyond, there is the possibility of some Hawthorne[25] effect of openly using audio- or video-recording equipment. An alternative to openly using audio- or video-recording devices would be to use them surreptitiously, without the awareness or consent of the people being studied, but this is generally considered unprofessional and unethical. There may, however, be scope to use recordings which have been made as part of everyday professional practice (e.g. in police interrogation rooms, or some psychiatric consulting rooms) if permission to use them for research purposes is subsequently sought.

Triangulation (see Section 8.13.2) is an important analytical technique of participant observation and ethnography.

8.9.4 Implications for government research, evaluation and policy making

Government researchers and policy makers may well ask what all this has to do with policy making, policy implementation and policy evaluation. There are a number of ways in which participant-observation and ethnographic inquiry can help develop, implement and evaluate policy. They can:

- Provide robust evidence on the processes by which front-line agencies work and how these might operate to promote successful implementation, or militate against successful implementation.
- Provide a 'window' through which the dynamics and decision-making processes of organisations can be observed first-hand.
- Identify variations in the social and cultural environment within which policies, programmes and projects are expected to work.
- Identify real-life drivers of policy success and failure.
- Identify real-life pinch points, barriers or vulnerability factors which might undermine the successful delivery of policies, programmes and projects.
- Identify key personnel who might operate as 'product champions' for policies, programmes and projects.

[25] Scriven (1991, p. 186) defines Hawthorne effects as: 'The tendency of a person or group being investigated, or experimented on, or evaluated, to react positively or negatively to the fact that they are being investigated/evaluated, and hence to perform better (or worse) than they would in the absence of the investigation, thereby making it difficult to identify any effects due to the treatment itself.'

- Identify key personnel, groups and values that might stand in the way of successful policy implementation and delivery.
- Identify process and outcome measures that have external and ecological validity (i.e. are meaningful and relevant to the communities that Government serves).
- Identify ways in which different policy-evaluation methods might be possible or unlikely to succeed.

8.10 Documentary research

8.10.1 What is it?

Most evaluations are likely to use documents as part of an evaluation. In some instances, documents are the source for 'reconstructing' a baseline for evaluation, recording what a project intended to achieve, the conditions it developed from, how the programme developed, what changes were made and how they were implemented. For example, documents can help us understand how professionals present themselves to service users, how they establish 'authority', and how this might make it difficult for a service user to make the most effective use of the service. In other instances, the documentary analysis accompanies other data collection methods, for example, in-depth interviews or surveys.

Most social researchers would agree that documentary data, like all other forms of data, are socially produced, that is, they are produced on 'the basis of certain ideas, theories, … principles,' and written for specific purposes and audiences, which in turn shaped their content and form (Macdonald and Tipton, 1993, p.188). Documents and records are never simple facts, but are mediated by the social context in which they were produced. Any documentary analysis needs to take this on board. The distinguishing feature of documents is that they have a historical dimension – that they are separate from the 'author', the 'producer' and 'user' by space and time. This may pose a specific challenge to the users of evaluation research, because there may be no running commentary available to provide additional information about the production and intentions of the documents. However, it means that documents can provide another 'reading' or perspective on an event or process, one that does not rely on a retelling or narration of it, with all the biases or inaccuracies that might bring.

Documents used in research and evaluation include public records, for example, legislation, parliamentary papers, administrative and historical records, annual and financial reports, minutes of meetings, strategy plans, policy papers; private papers, such as letters, diaries and notebooks; and other, non-text documents such as photographs, maps and plans, and

even buildings (Mason, 2002)[26]. They can also include media sources, such as newspapers, magazines and leaflets. As new technologies are proliferating, new forms of public and private documents are developing. They include video and DVD records, records resulting from the use of new web technology (Internet discussion sites, interactive websites, etc.), all of which have the potential to make a valuable contribution to evaluation research. This section will limit itself to discussing text-based sources.

One way of distinguishing documents is by asking who produced them, why and for whom. Ricoeur points out that all texts are written 'to do something' (Ricoeur, 1996). Another approach involves analysing the 'type' of record under investigation. Lincoln and Guba (1985) distinguish between 'records' where formal transactions are documented (for example, driving licenses document the passing of a driving test), and 'documents', which are produced for more personal reasons, for example, memoranda or notes. Transactions and records offer different types of information. Others have categorised textual documents according to the their closeness to speech. E-mails used in an organisation, for example, would be closer to speech than Acts of Parliament, which are stylised in form and content and utilise a 'full state technology of power' (Hodder, 2000, p. 703).

Evaluators are not restricted to using existing documents; documents can be produced as part of an evaluation exercise. For example, documentary evidence can be solicited by asking participants to keep diaries in order to produce time series of events. Still, the same questions about the conditions of its production apply, albeit in the context of the evaluation itself.

8.10.2 How are they used?

Documentary analysis can make a contribution to both formative and summative evaluation. In summative evaluations, documentary analysis can provide evidence on, for example, whether a policy has achieved its objectives and what the changes are. In formative evaluation, documents can provide evidence on the process for instance, of how a new service is developing over time; highlighting where barriers continue to be encountered. Other uses include the construction of different/alternative perspectives to those produced through other data source. Hammersley and Atkinson, for example, talk of documentary analysis as 'giving voice' to muted and suppressed groups, which might otherwise not be heard in the course of the evaluation exercise (1995). Two ways in which documentary research is used:

- as an independent dataset

[26] The use of documents in social research and evaluation has a long tradition, going back to the earliest days of social science research. Founders of social science methods, such as Max Weber in his study of bureaucracy and Emile Durkheim, in his study on suicide are based on the analysis of public documents (Macdonald and Tipton, 1993). And the Chicago School used documentary analysis in some of its classic studies (Hammersley and Atkinson, 1995).

  - Documents can be a historical 'audit', providing or supplementing information that would otherwise not be traceable; for example, it could be used to fill gaps when central stakeholders or staff cannot be interviewed as part of the evaluation, either because there are issues of confidentiality or because they have left a service.
  - Documentary analysis can track processes over a period of time (for example, the way a policy process emerged over time).
  - Documents can 'unlock' the otherwise 'hidden' history of a process or organisation and fill in 'gaps' in personal memories of interviewees and in what they are be prepared to say.

- as part of a larger research design
  - Documentary analysis can inform the design of research questions and other data collection methods (for example: carefully chosen documents could be used in focus groups to explore an issue further; findings of documentary analysis can inform the questionnaires and topic guides; a written record of the policy objectives can be used to further probe interview participants or focus groups; a leaflet explaining the key components of a service could suggest additional lines of inquiry in a survey).
  - Documentary analysis could be used to look at differences between documents and between documents and interview accounts in order to explain how differences in perceptions of a service can arise.

- It can be used for triangulation (see Section 8.13.2).

Example 8.3 Hospital Trust Mergers in London (Fulop et al., 2002)

This research project studied the processes and outcomes of mergers and reconfigurations of hospital trusts in London between April 1998 and 1999. The study employed a mixed-method approach. It included a management cost analysis, a cross sectional study of nine mergers between 1998 and 1999, and six case studies of trust mergers at two intervals. The findings of the study assessed the outcomes of hospital merger in terms of service delivery; whether trust size matters; on management structures and organisational culture; and on staffing and financial issues.

Documentary analysis formed an important dimension of the research:
In the cross sectional study, public consultation documents, outlining the reasons for the proposed mergers, were analysed to identify drivers informing the merger plans and the favoured organisational structure. Drivers identified included the need to make savings in management costs and invest them into patient services, to safeguard and develop specialist clinical services and units and to address staff retention and advancement issues.

Findings from the documentary analysis were used to inform interviews with key informants in the cross sectional study and in the case studies. For example, in the interviews with key informants views on the relative importance of stated drivers in the merger process were explored; informants were asked about additional reasons for the merger (unstated drivers) and whether objectives had been achieved or were likely to be achieved.
http://bmj.bmjjournals.com/cgi/content/abridged/325/7358/246

Different techniques in documentary analysis have been developed and continue to develop, drawing on different disciplines and expertise, for example, on literary, reflexive and interpretative techniques (Mason, 2002). They include:

- *Theoretical analysis,* where documents are studied to as to whether they support a pre-defined theory or explanation. For example, are the stakeholders' theories about a chain of events or change in a service reflected in the paper trail of memos, reports, etc.? Do the documents reveal other factors that need further exploration?

- *Structural analysis*, where the structure of the document is studied in detail, i.e. how it is constructed; which context it is set in; how it conveys its messages. For example, what can the structure, content and format of a particular health information leaflet tell us about the type of message, the target group and the expectations about behaviour change? Does this match what interviewees have told us about these issues in interviews? What do users of the service make of the leaflet?

- *Content analysis*, where the information in particular types of documents is studied and compared. For example, evaluators

could compare tender, contract and training documents to 'chart' the changes in thinking about the service from its initial inception to its implementation and to identify significant turning points in the process of developing the service. This could then be used to inform the questions for stakeholder interviews, develop the sample of stakeholders and identify further areas of investigation.

Less technically, documents may be analysed to bring context to the evaluation and to triangulate findings with other data sources (see Section 8.13.2). It would be advisable to seek expert advice on the specific requirements and skills for each technique that is to be used.

8.10.3 Implications for government research, evaluation and policy making

Being clear about what the document can and cannot deliver, what role it will have in the evaluation and how the claims made about a document can be substantiated will have a bearing on the inclusion of documentary analysis in the evaluation. Studies that use documentary analysis should have a clear description of:

- which documents were used and why;
- how they were analysed; i.e. the study should present an analysis framework for the documentary analysis;
- how the documentary analysis relates to other aspects of the evaluation.

One of the problems for including documentary analysis in the research is access to documents. Not all documents that are required may be readily available – many government documents are restricted; documents may be lost. The use of private documents obviously raises issues of confidentiality.

8.11 Conversation analysis and discourse analysis

8.11.1 What are they?

Conversation analysis and discourse analysis place a particular emphasis on talk and text as data sources. They focus on how the spoken and written word are used in everyday settings and how they structure communication and understanding. Discourse analysis examines the ways in which social events are reported and represented by people with different interests, and it uses this to identify the underlying values and agendas of different social groups.

8.11.2 How are they used?

Conversation analysis studies naturally occurring talk in order to identify the underlying rules of social interaction and what makes everyday social life ordered.

This requires data-gathering methods using audio- and video-recording, detailed note-taking and textual sources such as official documents, policy statements, briefings and media articles. These 'naturally occurring' data are then subjected to detailed analysis and interpretation using turn-taking analysis (the detailed analysis of a dialogue between two or more speakers), speech-act analysis (the study of meanings, structure and symbolism embedded in speech and dialogue) and other sociolinguistic methods.

8.11.3 Implications for government research, evaluation and policy making

While conversation analysis and discourse analysis may not initially be seen as a method of policy evaluation, they have been used by social researchers to study the ways in which politicians speak and the rhetorical devices they use to influence people's understanding of politics, policy and society (Atkinson, 1984; Brown, 1987; Wilson, 1990; Schaffner, 1997). They have also been used to help identify implicit assumptions and agendas of policy makers and of policy statements. Such approaches may have much to offer qualitative methods of policy evaluation, especially at a time of wide suspicion and concern about political 'spin' and the manipulation of social and political reality. They are important in understanding how professionals/staff interact with service users. For example, they can be used in looking at how work-focused interviews are conducted – who raises the issue of work under what circumstances, how far different staff challenge people's stated reasons for not being able to work, which ways of doing this are most effective/least threatening. Robert Dingwall (1997) has used it to look at how far mediators in family breakdown really are unbiased or neutral.

8.12 Biographical approaches, life histories and narratives

8.12.1 What are they?

This group of approaches involves collecting detailed descriptions of a person's whole life, or of a key event or period within it. They use a particularly intensive form of data collection and analysis. For example, researchers will often go back to an informant repeatedly to collect further information or conduct a series of interviews. Or they may follow an individual over a period of time, collecting detailed data on developments in their lives. They may study, for example, the perspective of a family by interviewing different family members, or the

members of specific communities. The approaches have been utilised, for example:

- to study the recovery from alcohol and drug dependency;
- to collect (auto)biographies in mining communities;
- to record narratives of organisations and of historical events and periods;
- to give (authentic) voice to and contextualise the experience of homelessness.

Biographical, life history and narrative approaches draw from a range of interpretivist schools in the social sciences. At their core lies an understanding of social actors as participants in their social environment. People are perceived as being able to make sense of and convey experiences through their accounts[27]. Personal accounts and shared narratives are a source for tapping into social patterns, rules, experiences and for exploring individual and collective meanings – they are another way into 'reading social, cultural and economic history' (Mason, 2002, p. 57). Biographical and narrative approaches are also able to engage with sensitive and emotional experiences.

8.12.2 How are they conducted?

Biographical, life history and narrative research can use accounts generated specifically for the research project, but existing accounts, such as verbal accounts, stories, written biographies and video accounts have also been used.

Techniques are based on in-depth interviewing (see Section 8.6). Some approaches have adapted methods used in psychotherapy and counselling to encourage people to speak about their experiences and lives (Wengraf, 2001). Some researchers will return repeatedly to informants to attempt to understand a specific narrative – turning the interview process into a dialogue between researcher and researched. Some approaches perceive of the interview process as following a dynamic process of narration and memory, which can be guided by the interviewer through the use of specific interactive techniques (Wengraf, 2001). In this version of narrative interviewing, highly structured techniques of analysis are employed. More indirect ways of collecting data include, for example, storytelling, and the keeping of diaries and other records by informants (for an example see also Schuller et al., Example 8.7).

[27] Narrative and biographical approaches have diverging views about how accounts relate to the actual lived experience and to what degree social actors understand their own complex life actions. For a discussion of this, see Wengraf, 2001.

8.12.3 Implications for government research, evaluation and policy making

Currently, biographical and narrative approaches seem to be underused in policy research. This is somewhat surprising, given that they are a unique source for gaining access to the processes and impacts of interventions at an individual level. Biographical and life history approaches can personalise and bring to life consequences of interventions. Narratives and biographies are also a means of fixing otherwise elusive processes, such as cultural changes. For example, narratives and biographical approaches contribute to the evaluation by:

- illustrating the personal consequences of an intervention;
- tracking the longer-term and complex outcomes of change at individual and group level;
- integrating the personal and emotional with the social and structural dimensions of peoples lives, values and perspectives;
- providing a voice to otherwise neglected perspectives and groups (Shaw, 1999).

For example, asking a disabled person to narrate and describe key events and periods of their life would help others understand to what extent previous and current interventions have shaped their life chances, their views and perspectives. They could help in examining the 'fit' between an intervention and the material, personal and emotional circumstances of a person's life, and also, through a personal 'narrative', the impact of an intervention could be exemplified. Biographical, life history and narrative approaches could make a critical contribution to an understanding of a fluid process such as social exclusion, because it would help to identify the periods in someone's life when they felt more or less included/excluded and to understand the structural and personal circumstances which were relevant at those times and which might have contributed.

The approaches can be used in a variety of ways, for example:

- as illustrative case studies, i.e. thoroughly researched 'individual accounts' showing the longer-term impact of an intervention;
- as 'histories' of individuals and groups;
- as particular themes within the evaluation;
- as test cases for findings reached.

The analysis and writing up of live histories, narratives and biographies tend to be time-consuming and complex. There is usually a great deal of redrafting, working with research participants and other researchers necessary to develop the material to its full potential. This in turn

requires a high degree of specialist skill and experience on the part of the analysts.

8.13 Design

Having described the key data collection methods used in qualitative research, we now go on to look at important aspects of the design, sampling, analysis and presentation of qualitative research.

8.13.1 Key principles

The key principles of qualitative research design share much with those relevant to quantitative research. In either case, there must be a clear set of research questions that are sufficiently specific to be actionable in the research. Research questions that are broad or vague can easily lead to unsatisfactory studies that simply do not produce new insights. More time spent early on sharpening up what needs to be known will help to prevent this. However, additional questions often arise during the research process, which means the initial questions may need to be revisited and amended in the light of these developments.

There needs to be coherence between the research questions and the populations studied. These should be the populations that are going to give the most direct and insightful information on the subject matter. There needs to be some thought as to which subsets of these populations are particularly critical to include, or which should be excluded. For example, does the evaluation require information from policy developers or from staff delivering the service and, if so, in what roles? From their managers? If so, at what levels? From stakeholders, from current participants, eligible non-participants, past participants, etc.? In each case, which types of members of these groups are likely to have different but important things to say?

Although it is unusual to have formal 'control' groups in qualitative research designs (since these are usually associated with the measurement of difference), building comparison into research designs can be very helpful and can lead to more in-depth understanding. So, for example, a study looking at a particular phenomenon among lone parents (such as attitudes to work) might be enhanced by including couple parents. Comparing the responses of the two groups will help with understanding what is a function of being a lone parent, as opposed to just being a parent.

There should also be coherence between the research questions and the settings studied. Qualitative fieldwork is usually focused on a small number of different sites or areas, carefully selected to provide coverage of different types of contexts. These might be different institutions, organisations, labour markets, or types of urban or rural setting, for example.

There should also be a logic between the research questions and the data collection methods used. For instance, are naturally occurring data needed because, say, what is being researched is best illuminated by observing behaviour or interaction, or by reviewing existing documents? Or is what is sought people's own accounts, explanations and interpretations, in which case should this be through individual interviews or group discussion?

And there needs to be a logic to the timing of episodes of data collection. This involves some careful thought about which perspective, or perspectives, on what is being researched will be most illuminating. Is it the early delivery of the service that should be researched, or later stages when delivery incorporates lessons learnt? Is it the experiences of participants as they first contact the service, or as they use it, or after they have stopped using it, that are needed? In practice, a thorough evaluation design will often use more than one point of data collection and will involve both repeat cross-sectional and panel designs:

- Repeat cross-sectional design is where different people are selected at different points in time. For example, one set of staff might be interviewed when implementation is being planned, another set in the early days when the service is 'bedding down', and a new set of people in later stages when relationships have been formed and ways of working established.

- In a panel design, the same people are interviewed more than once. For example, clients might be interviewed in the early days of their use of the service, later while they are still in contact, and then some time later when they have left the service but when more than the immediate impacts have been experienced.

The feasibility and appropriateness of a proposed design and approach within the actual research setting – the service or policy being evaluated – is of critical importance. Where what is being researched is relatively new or has not been the subject of much scrutiny, early field visits and discussions with those involved in the design, delivery and use of a service can be invaluable in informing the research design.

A further issue of key importance in evaluation is to give early thought to the criteria against which a policy or service is to be evaluated, and how these are to be generated. Qualitative research provides an opportunity to generate evaluative criteria from the viewpoint of clients, staff or stakeholders, as well as using criteria which are established *a priori* by policy makers.

One of the key advantages of qualitative research is that it is flexible and can, more easily than quantitative research, be adapted. As more is learnt about the research phenomenon and setting, it is not uncommon for there to be a need to modify the study sample, the research

questions being asked and how they are formulated in the study, or the fieldwork approaches. Changes might also be made because the information requirements of Government change – new issues are thrown up by what is learnt early on, or by changes in the broader policy context. Although this is not an excuse for poor planning, the process of design in qualitative evaluation is a continuous one and the scope for modification an advantage.

8.13.2 Triangulation

A good research design will often include elements of triangulation. Triangulation means bringing together different types of data, or sometimes different ways of looking at data, to answer the research questions.

Denzin (1989) describes four types of triangulation: *methodological triangulation, data triangulation, investigator triangulation and theory triangulation*, of which the first two are most commonly used in government evaluations.

- Methodological triangulation means combining different research methods. Denzin distinguishes between 'within method' triangulation (where, for example, a range of different lines of questioning might be used to approach the same issue), and 'between method' triangulation (where different data collection methods are combined). For example, observation of client–staff interactions might be combined with analysis of documents (e.g. case notes or guidance notes), or with interviews (e.g. to ask the participants about their experience of the interaction), or with group discussions (e.g. to ask staff how they vary their approach in different cases). Since each of these methods has its own strengths and weaknesses, combining them provides a more rounded picture of what is being researched.

Methodological triangulation might also involve combining qualitative with quantitative data. Some people would criticise this approach on the grounds that the fundamental philosophical beliefs that underpin the two methods are so different that they can never be combined meaningfully. It is certainly true that qualitative and quantitative research address different research questions and generate very different types of data. However, many evaluation designs combine the methods to very useful effect. Qualitative research can be used:

- *before* quantitative research, for example, to map and clarify issues for coverage, generate hypotheses and define terminology and concepts;
- *concurrently*, to explore in depth issues measured by quantitative research;

- *after*, to provide further explanation, particularly of unexpected findings for research among key subgroups identified by quantitative research.

For example, evaluation designs typically use surveys to measure levels of satisfaction with the service and outcomes for different groups, and qualitative research to look at how the service is delivered and experienced, and to understand what works for individual people in different personal circumstances.

- Data triangulation means combining data from more than one source, say, from a number of settings, points in time or types or groups of people. For example, a range of different service delivery locations might be selected or the study might be conducted with staff, clients and non-users of the service to explore differences in their experiences and perceptions.

- Investigator or analyst triangulation is slightly different. It involves more than one researcher looking at the data so that they can either check or challenge each other's interpretation or deliberately approach the data from different angles.

- Theory triangulation means looking at the data from different theoretical positions in order to explore the fit of different theories to the data and to understand how looking at the data from different assumptions affects how it is interpreted. In the evaluation context, for example, it might involve looking at data from a goals-based and a goals-free perspective to understand the differences in how it meets those different sets of criteria (see Background paper 1: what is policy evaluation?).

8.13.3 Case studies

Method and data triangulation come together in case studies. The term 'case studies' is used in different ways, sometimes to imply a focus on a single case, but in this chapter we use it to mean bringing together different perspectives to understand a context, or a set of contexts, in more detail (Patton, 2002; Robson, 2002; Yin, 1993 and 1994). The different perspectives might come from the use of qualitative and quantitative research, from the use of different qualitative methods, or from drawing together the accounts of different players or groups – such as staff, managers, clients and other stakeholders. The context might be:

- a relationship – between a couple, for example, or a professional and their client;

- an organisational entity – such as a local education authority, or a school, or a class;

- a process – for example, the legal resolution of relationship breakdown, or the determination of an application for a social security benefit.

Case study designs are used where what is required is very detailed in-depth understanding that is holistic, comprehensive and contextualised. They allow comparisons to be made between different actors within a single case, between cases, and between groups across cases. So, for example, in the context of schools-based research, one could look at how different people within a school have different understandings of a new educational initiative, or at how different schools have implemented it differently, or at how head teachers view the initiative compared with teachers or pupils.

Case study designs are intensive and thus can be expensive, time-consuming and complex, but they can bring very powerful understanding to policy evaluation. Observational and ethnographic data can be triangulated by other people who took part in the observed activity, by other researchers (especially where audio- or video-recordings can retrieve what took place), and by documentary analysis (e.g. of case notes, office files, report, minutes, memoranda).

8.13.4 Making sense of triangulated data

There is some debate about whether the purpose of triangulation is to verify (that is, establish the truth) or to broaden and deepen understanding. The idea of using different approaches to verify understanding is increasingly challenged, particularly on the grounds that there is no single version of 'truth' that can be captured, and that different methods or approaches inevitably produce different types of data which are unlikely to be concordant (Brannen, 1992; Denzin, 1989, 1997; Fielding and Fielding, 1986; Flick, 1992; Hammersley and Atkinson, 1995).

The purpose of triangulation is more often understand to be to add richness, depth and breadth to a study. Different methods, or different populations, are unlikely to yield data that calibrate precisely. But drawing together different perspectives and types of information provides a more rounded understanding. Where inconsistencies are found, they may be a key finding in themselves, for example, highlighting that staff and clients have different understandings of why people use the service, or that what staff describe as 'client-focused and responsive' practice is not experienced as such by clients. Or they may prompt further examination of datasets to see if explanations for the inconsistency can be found. But the approach is not generally used to mediate between different methods or sources and to say which is 'right'. This type of in-depth analysis requires iteration and time, and the value of a study using triangulation can be lost if time is not allowed for it.

Example 8.4 Evaluation of New Deal for Lone Parents (Lewis *et al.*, 2000)

This study was commissioned by DWP as part of the evaluation of New Deal for Lone Parents in its earlier stages. The study used in-depth interviews with lone parents who had used NDLP, followed by focus groups with personal advisers who deliver the service. The qualitative research with lone parents showed how their work experiences and broader aspects of their lives gave rise to very different needs of the service. It also highlighted diversity in their experiences of the service in terms of the intensity of contact, the breadth and depth of support given, the allocation of responsibility for action, the pace and degree of work-focus, the amount of personal support and attention to underlying issues and whether or not lone parents found work after using it. These differences were not fully explained by lone parents' different circumstances.

The research with personal advisers used vignettes – short examples of cases – drawn from the lone parents interviewed and more general discussion of approaches and practices. This highlighted, and provided reasons for, differences in the way in which personal advisers approach their jobs, which appeared to be important explanation for the diversity of experiences found among lone parents. The research highlighted that these differences in personal advisers' practices were influenced by a range of factors, which shaped advisors' perceptions of the needs of clients, the scope of the service to meet them, and views about appropriate ways of working in different cases.

http://www.dwp.gov.uk/asd/asd5/rport122/main.pdf

Example 8.5 Experiences of and Attitudes to Disability (Grewal *et al.*, 2002)

This study, funded by DWP, involved in-depth interviews and focus groups with disabled people, focus groups with non-disabled people and a large-scale survey of disabled and non-disabled people. The qualitative work preceded the survey and was used to shape the survey coverage and develop specific questions.

For example, the qualitative research showed that very different meanings are attached to the concepts of 'prejudice' and 'discrimination'. The former is seen as an expression of perceptions of difference, sometimes unintentional, such as staring or asking inappropriate questions. The latter is seen as more serious, involving organisations or their representatives preventing someone from having an equal role in society. The qualitative research highlighted many examples of each in the experiences of disabled people, and provided understanding of their impacts, their perceived causes, and views about what might be done to tackle them. The survey was then used to look at the extent to which disabled people had encountered different forms of discrimination and prejudice, deriving examples from the accounts of those who had participated in the qualitative research.

http://www.dwp.gov.uk/asd/asd5/rrep173.asp

8.13.5 Validation

The concept of validation is sometimes linked with triangulation. Validation means bringing to bear on research findings other forms of knowledge or information. Again, there is some debate about whether this is to verify the findings (i.e. to establish their 'truth' or their credibility) or more generally to bring another perspective. The most common forms of validation are:

- *participant or member validation*, which involves taking findings back to research participants to see whether the researcher's interpretation of data accords with the perspectives and intentions of participants. This can be a useful check on bias and the quality of research, although, of course, research participants may not necessarily have a neutral or objective take on their own behaviour. It is also a chance to fill gaps in data or in explanations, and it is an important aspect of participatory and emancipatory evaluation.
- *peer or expert validation*, which involves taking findings to a wider group who know the research phenomenon or

population well, to see how the findings chime with their knowledge and experiences.

- *validation through corroboration* with other research, where other research is used to help to assess the credibility of the findings. Again, neither set of findings can necessarily be privileged, but one would expect to find robust explanations for and checks on findings which were out of line with other research.

8.13.6 Implications for government research, evaluation and policy making

Qualitative research design is a complex art, and it is a process – a series of decisions which need to be reviewed and revisited – rather than a one-off event. For government evaluators it is critical that:

- Research questions are clear, sharp and understood.
- There is a coherence between the questions and the research design in terms of the populations, settings, data collection methods and timing of data collection.
- Thought has been given to whether comparison needs to be built into the design to sharpen its focus.
- Different data methods and sources are used in combination to maximise the potential to provide full, insightful and credible answers to the research questions.

8.14 Sampling in qualitative research

8.14.1 Key principles

Qualitative research sampling has a quite different logic from that of quantitative research. The objective is to select the individual cases (which might be people, documents, visual images, events, settings, etc.) that will provide the most illuminating and useful data addressing the research questions. The intention is therefore not to provide a precise statistical representation of the researched population, but to reflect aspects of its diversity that are expected to generate insight.

There are two main approaches to sampling. In *purposive sampling* (Arber, 2002; Patton, 2002; Ritchie and Lewis, 2003) sample cases are chosen deliberately to represent characteristics known or suspected to be of key relevance to the research questions. The selection criteria are prescribed at the first stage of sample design, based on a review of existing research or information, discussions with people with expertise or experience in the research area, or on the researcher's hypotheses. The required composition and size of the sample is then determined and individual cases selected to fit the required composition. Purposive

sampling is also called 'judgement sampling' (Burgess, 1984; Honigmann, 1982) or 'criterion sampling' (LeCompte and Preissle, 1993).

In *theoretical sampling* (initially Glaser and Strauss, 1967; Strauss, 1987; Strauss and Corbin, 1998; see also Bryman, 2001; Finch and Mason, 1990; Seale, 1999), the researcher makes decisions, as the study proceeds, on the basis of emergent theory from their analysis of initial data, about the type of data to collect and participants to involve next. Theoretical sampling is strongly associated with grounded theory – an approach that focuses on generating theory inductively from data. Here, the sample evolves during the course of the study as theoretical thinking develops. The process is an iterative one: an initial sample is selected, the data analysed and an emergent theory developed; a further sample is then selected to help refine an aspect of the emerging theory and the data analysed; this leads to further development of theory, more case selection and fieldwork, and so on. The process continues until 'data saturation' is reached – the point where it is judged that no new insights would be obtained from further cases.

Theoretical sampling is likely to be more appropriate where theory development is a central objective of the study, or for a research subject or setting about which too little is known to allow selection criteria to be prescribed before sampling and fieldwork begin. However, there is much less clarity about funding and time requirements than with purposive sampling, and purposive sampling is more often used in government evaluations, although with the flexibility to modify sampling and data collection as appropriate.

It is important to understand that the deliberate, non-random selection of cases in qualitative research is its strength, not a weakness. Sample design and selection is systematic and theoretically based, but follows a different logic to quantitative sampling. Some data collection methods and some circumstances require more informal approaches. For example, in ethnographic and participant-observation research there will be less scope to select cases for inclusion at a distance from the research setting and a greater need to take advantage of opportunities as they emerge. Nevertheless, the researcher is making rational and defensible decisions that have a coherence with the research questions.

Qualitative samples need to be large enough to include key subgroups and to reflect diversity. The emphasis is on mapping and understanding issues, rather than counting or numerical representativeness. In fact, large samples are a positive hindrance. The data generated in qualitative research are rich and intensive. Depth lies in the quality of data collection and analysis, not in quantity. The appropriate size of a sample will vary and is always a matter for judgement, but it also needs to be reviewed during fieldwork and as fieldwork draws to a close so that gaps in sample coverage can be filled. The same principles apply for group data collection methods, such as focus groups.

Finally, the sample frames used in qualitative research are varied, as in quantitative research. Broadly, they may be existing data sources, such as survey samples, administrative records, registers or databases, or sources which are generated specifically for the research.

This latter group would include, for example, household screens (where households are approached, usually without prior selection, and a short questionnaire used to ascertain whether the households include anyone eligible to participate). Networking through organisations is also sometimes necessary to generate a sample, where organisations are asked to put potential participants in touch with the research team. For some studies it would be appropriate to use a flow population as a sample source: approaching people in a particular location or setting that is relevant to the research, such as a GP's waiting room or a Jobcentre Plus office. Sometimes the only feasible approach is to use snowballing, where research participants are asked to introduce the research team to another person in the required study group – a method primarily used to find small and dispersed populations where the relevant sample characteristics are unlikely to be widely known to others outside the group (classically, sexuality).

8.14.2 Implications for government research, evaluation and policy making

In assessing the quality of sampling in a qualitative research study, government researchers and policy makers will want to be concerned with:

- whether the study group has been defined and classified in a way that is meaningful, given the diversity present in the relevant population;
- whether the rationale for case selection is logical and likely to include all key subgroups;
- whether the sample size is sufficient to map the required sample with adequate depth in key subgroups, but not so large as to overwhelm the research team or be unfeasible with regard to the time and resources available;
- whether the sample frame used is inclusive and comprehensive and does not build bias or gaps into the sample.

8.15 Analysis of qualitative data

8.15.1 Key principles

Approaches to the analysis of qualitative data vary, particularly reflecting different assumptions about knowledge, social existence and the nature of qualitative research. Some distinctive theoretical approaches to qualitative research offer frameworks for analyzing qualitative data, such

as phenomenological analysis, grounded theory, qualitative comparative analysis and analytic induction (Patton, 2002). More generally, approaches to analysis vary, for example, according to:

- the status of the data – whether it is treated as a representation of 'reality' or as social constructions;
- how the data are condensed or reduced;
- how concepts are derived from or applied to data, and the level of abstraction involved;
- the extent to which data are retained in context;
- the way analysed data are accessed and displayed;
- the place of the researcher in the analytical account.

In thinking about analysis, it is helpful to distinguish between data management (when data are labelled, ordered or summarised) and data interpretation (which involves conveying or displaying data and finding meaning within data).

One distinction between methods for data management is whether they are paper-based or computer-assisted. Paper-based methods differ in terms of whether they keep cases separate or bring them together in a single framework, and whether they involve working with raw data or summarising data. They include:

- case summaries in which individual summaries are constructed for each case;
- thematic data ordering in which raw data from different cases are brought together under thematic headings;
- matrix methods within which each individual case is summarised thematically within a single framework;
- mapping methods where thematic or cognitive maps are designed based on constructions and linkages within the data.

Computer-assisted methods for qualitative data analysis (of which NUD*IST, Ethnograph, NVivo, winMAX and ATLAS/ti are the main packages) vary in terms of: how data are entered and stored; approaches to coding; data-linking mechanisms; mechanisms for search, retrieval and display; mechanisms for attaching memos or notes to codes; and mechanisms for tracking and recording the analysis approach (Fielding and Lee, 1998). A distinction is also drawn between text retrieval, code and retrieve approaches, and theory-building or conceptual network building, although the more popular packages fulfil most or all of these functions.

There are no rules about whether or when paper-based or computer-assisted methods are superior, or which approaches within each are to

be preferred. What is more important is that the selected approach is appropriate to the key features and objectives of qualitative research and will aid the later stage of interpretation. Important considerations in the data management process are (Spencer *et al*., 2004):

- remaining grounded in the data – using a structure that allows analytical ideas and concepts to emerge from data rather imposing them;
- capturing the synthesis of data – so that the link between raw and summarised data is preserved and can be reviewed;
- permitting searches within cases and between cases – so that the integrity of individual cases is retained as well as facilitating thematic analysis and comparisons between cases;
- systematic and comprehensive coverage of the dataset – rather than a selective or biased focus on themes, issues or cases;
- flexibility – so that the structure within which data is managed can be modified if necessary;
- transparency – so that the approach can be described and made accessible to others.

Whatever the data management method used, the process of interpretation is an intellectual one in which the researcher must draw on their own cognitive and conceptual skills. It involves 'creativity, intellectual discipline, analytical rigor, and a great deal of hard work' (Patton, 2002, p. 442).

The outputs of this involve *descriptive analyses*, the identification of patterns and associations in data, and interpretive and explanatory accounts. In descriptive analyses, the researcher is concerned with the substantive content of their data. They seek to identify key dimensions of phenomena, experiences, attitudes or behaviours, and to construct, categorise and display them in ways that illuminate the data. In speech- and text-based methods they are concerned with the language used by respondents, in observation with their detailed behaviours, and they seek to display the richness, colour and texture of the original data. Lofland (1971, p. 17) describes this as 'documenting in loving detail the range of things that exist'. In evaluative research, this might involve describing, for example, the way in which a service is organised and delivered; the types of support provided; the circumstances of service users; or the range of (intended and unintended) outcomes experienced.

In *associative analysis,* the researcher looks for patterns, replication and linkages in the dataset. These might be associations within the data, such as linkages between attitudes, or between attitudes and behaviours, or between circumstances and needs. Or they might be patterns in the location of a phenomenon within the data (which types of people held a particular view, for example), or differences in how it is manifested

among different cases. The purpose here is not to display differences or associations quantitatively, but to use the associations or patterns found in the data to enrich understanding of the phenomenon in question, and to prompt further searching of the dataset to understand why the association or pattern exists. In evaluative research, this might involve, for example, looking at which providers deliver the service in different ways; which service users experience positive outcomes; linkages between service delivery and outcomes; or how requirements of the service are influenced by circumstances.

In *interpretive and explanatory analysis*, the researcher builds explanation from or finds explanation within the data for the views, behaviours or accounts described, and for the associations and patterns found. These are not narrow deterministic explanations. They are sometimes described as explanations at the level of meaning rather than at the level of cause (Hughes and Sharrock, 1997), that is, explanations that reflect the social construction of phenomena and the meaning attributed to them by research participants. Where 'causes' are offered, they are not based on mechanistic linkages between isolated variables, but on an examination of the way in which different meanings and understandings come together to influence an outcome, behaviour or decision (Patton, 2002), and on identifying the nature of, and relationships between, different influences or contributory factors. They might be, for example, explanations for why a service is delivered differently by different providers; why some users experience positive outcomes and others do not; how perverse or unwanted consequences arise; or how and why the needs of different service users vary.

The 'evidence' on which explanations are based may be (Ritchie et al, 2003):

- explicit statements made within an individual account;
- constructed by the researcher, based on an underlying logic inferred from the data – this may be underpinned, for example, by the juxtaposition or interweaving of themes, the absence of something in one case which is found in another, or the repeated coexistence of phenomena;
- everyday or 'common sense' assumptions;
- explanatory meta-concepts developed from the study;
- concepts or explanations drawn from other studies;
- existing established theoretical or conceptual frameworks.

For all three types of analysis – descriptive, associative and explanatory – the process involves reviewing and interrogating the dataset. This interrogation will be prompted by the research questions, by the researcher's own hunches and by other research or theoretical frameworks. It involves moving between different levels of abstraction

and conceptualisation and carrying out both detailed study within individual cases and comparisons between cases. The search is systematic and comprehensive and the researcher looks for convergence and divergence between cases, for typical and outlying cases, always testing the fit of an association or explanation across the dataset, expecting to find multiplicity and looking for rival explanations.

The quality of qualitative analysis comes from the creativity, flair and insight of the questions the researcher asks of the dataset. It is underpinned too by the analytical conscience they show in systematically reviewing the fit of apparent findings across the dataset, looking for diversity and multiplicity and refusing to be content with single stories or explanations. Qualitative analysis requires considerable intellectual transformative work with the data, which is why some commentators have referred to qualitative analysis as being akin to the artistic process. It is certainly a time-consuming activity and should be allocated enough space in the overall research process.

Example 8.6 A qualitative study of mentoring interventions with young people (Philip *et al.*, 2004)

This study, funded by the Joseph Rowntree Foundation, explored how young people interpret mentoring in relation to other social and professional relationships, and how their own backgrounds and experiences interact with mentoring. It involved a literature review, analysis of documents, observation of mentoring relationships, two phases of interviews with young people in mentoring relationships and interviews with mentors and other stakeholders. The researchers make clear that they draw particularly on the research with young people to capture the transitions and changes in young people's lives which form the context of mentoring. The use of in-depth interviews in this study seems particularly suitable in addressing subtle and changing relationships. Early in the report individual in-depth examples are used to give a rounded and holistic picture of their lives and of mentoring within them.

The report looks at the underlying processes involved in mentoring – how mentoring relationships become significant, how they can be used by young people to test out ideas and identities, and the negotiations and processes involved in the mentoring relationship. The report also looks at how young people use and interpret mentoring, for example how they use skills they see as acquired through mentoring to negotiate with parents and friends; how they see them helping in rebuilding relationships; and how they help in reflecting on significant life events.

http://www.jrf.org.uk/knowledge/findings/socialpolicy/324.asp

Example 8.7 Learning in Adult Life (Schuller *et al.*, 2002)

This research study, conducted by a centre funded by the Department for Education and Skills, involved over 140 biographical interviews with adults involved in a range of different learning contexts. The research explored interactions between learning and life, and the effects of learning on people's lives.

The study explores these issues in particular depth. They look at how learning shapes identities which link with civically aware behaviour, and at the influence of personal and external contexts on this process. Individual biographies are used as examples to illustrate people's pathways into and around learning and the interactions between health, family and civil participation effects.

For example, in highlighting that a growth in self-confidence was the most fundamental and pervasive positive impact, the researchers identified 15 different benefits which respondents described as flowing from increased self-confidence. The analysis also explored in detail the direct and indirect effects of learning on social capital and social cohesion: impacts included the acquisition of civic competences by individuals and providing opportunities for civic engagement. The research analysed the impact on learning on social networks and the flow of social capital between groups, identifying four mechanisms of transmission.

http://www.learningbenefits.net/Publications/ResReplntros/ResRep3intro.htm

8.15.2 Implications for government research, evaluation and policy making

There are no standardised procedures for qualitative data analysis, but this does not mean that analysis should not be systematic, grounded in the data and defensible. Flair and creativity are undoubtedly important aspects of qualitative data analysis – and indeed of quantitative data analysis too. But the analysis process needs to be rigorous, comprehensive and involve very close attention to detail. Government researchers and users of research should expect to see:

- transparency about how analysis has been carried out, that is, in what form the data was used, how analytical categories were constructed and applied, how the data were searched and reviewed;
- analytical outputs that provide an accessible, structured and nuanced window on the data and which are oriented clearly to the research questions;
- multiplicity and diversity – it should be clear that the researchers have not stopped at the first and most obvious explanations or the most recurrent views or stories. For every finding, they should expect to see discussion of atypical or

divergent cases and evidence that the researchers have looked for alternative explanations.

8.16 Presenting and using qualitative research

Qualitative research evaluations – like those using experimental or quantitative methods – will rarely point directly to a single course of policy action. This is partly because a good piece of qualitative research will make a research user more, rather than less, aware of the complexity and diversity of their subject; and partly because good policy development often involves drawing on more than one piece of research or information. The worlds of research and policy making overlap but they are not identical in terms of the expertise and knowledge they draw on. Developing policy out of research is a creative and reflective process of translation and not simply a matter of direct application.

Nevertheless, there clearly are circumstances which should make research users in Government more or less secure about using a piece of qualitative research in the formulation and development of policy and in making decisions about moving from piloting to rolling out policies. Central here are the robustness and credibility of the research.

Reports of qualitative research evaluations should provide a detailed explanation of the design and conduct of the research, outlining key decisions made about: the choice of research methods; the selection of sites, populations and samples; the design and focus of data collection instruments; the fieldwork strategy and how it was implemented; and the approach to analysis. Reflections on limitations that flow from the design and conduct or caveats in using the research also add credibility and reassurance.

We looked in the previous section at features of analysis that lend credibility, particularly whether it is clear that the data have been explored systematically and comprehensively. Triangulation, validation and looking at how the findings relate to existing research can also be helpful.

In considering the findings, users should also look for evidence of a clear link between data and conclusions, that is, the building blocks or conceptual steps which led from data to conclusion should be displayed. Research reports should clearly address the original research questions and demonstrate how and to what extent they were answered in the research. The report should convey the depth and richness of the data. The reader should feel they have stepped into the research participants' shoes and that they have been given a chance to understand their world from within. The report should convey the complexity and subtlety of the research phenomenon, but in a way that is structured and makes it more, not less, clear. In evaluations, it should also be clear what criteria have been used in the appraisal of the policy or service and where these came from.

Determining what reliance can be placed on the findings and how they can be used also raises the issue of whether and how the findings can be generalised. i.e. whether they have a wider relevance beyond the specific study sample. The considerations involved will be slightly different depending on the relationship between what has been researched and the setting to which the findings are to be applied – whether they are the same setting, or whether what is envisaged is the application of findings from one setting to another. These require what are essentially different forms of generalisation, the former being *representational generalisation* (generalising from the study sample to the population it represents) and the latter *inferential generalisation* (extrapolating from the study setting or population to another setting or population) (Ritchie and Lewis, 2003).

In *representational generalisation*, the concern will be with how well the setting and population were represented in the research study design and sample, the quality of data collection and the credibility of the claims made. In inferential generalisation, an additional consideration will be the degree of congruence between the 'sending' context from which the findings derive, and the 'receiving' context to which their application is being considered (Lincoln and Guba, 1985). Assessing this requires 'thick description' (Geertz, 1993) of research settings and observations, so that they can be understood in depth by the research user, and similarities with and differences from the context to which the findings are to be applied can be understood.

Assessing how a piece of qualitative research can be used involves forming a view about its quality. Good quality standards and criteria for research are as important for qualitative research as they are for quantitative investigations. Evaluators, commissioners, policy makers and funders need to be able to make assessments about the quality of research studies they are using or intend to use. For example, research commissioners need to assess the quality of a proposal; policy makers need to make an assessment of the quality of research findings to include them in the development of new policies. Qualitative research studies have to be of a high enough standard to instill trust in the findings they develop, the arguments they present and the case they are trying to make.

Qualitative research has used peer review as a means of assessing quality – that is, research is assessed by others working in the field – for example, in appraising journal articles and in proposal assessments. While it is an accepted mode of quality assessment, it has its pitfalls in that it is often applied inconsistently and inadequately (Boaz and Ashby, 2001)[28]. As qualitative research has expanded and found entry into different areas of policy evaluation, the notion of a quality framework of qualitative research has been discussed as a means of promoting high-quality research standards in qualitative research. A number of quality

[28] Of course, this is also true of peer review in quantitative research.

criteria and frameworks of assessing quality have been developed in different research fields (Spencer *et al*., 2004).

However, the development of quality standards is not universally accepted and remains a contested area. Critics have pointed out that the diversity of philosophical and theoretical positions informing qualitative research makes it difficult to derive at common standards of quality. They point to different traditions of design, data collection, analysis and reporting/writing-up styles, which are difficult to bring under one common understanding of quality. Some commentators reject the idea of common quality standards for political reasons: they feel that the diversity of the field is a strength, and that standards would bring a push towards uniformity and would lead to preferential use of some techniques over others. Against this, the movement towards setting criteria for qualitative research argues that despite the great diversity and multiplicity of qualitative approaches, qualitative research and evaluation adheres to some core of quality principles. Developing frameworks injects openness, clarity and transparency into the process. Used with care, quality frameworks can champion innovative and diverse qualitative research approaches.

Quality in Qualitative Evaluation: A framework for assessing research evidence has attempted to stir a middle pass between developing quality criteria that qualitative research should adhere to and encompassing a wide range of qualitative research. It is based on four guiding principles for assessing quality, addressed through a total of 18 appraisal questions, which reflect the key features and processes involved in qualitative inquiry (Spencer *et al*., 2004, p. 20).

The guiding principles are:

- Research should contribute to knowledge and understanding.
- Research should be defensible in design.
- Research should be rigorous in conduct.
- Research should be credible in claim.

The 18 appraisal questions cover findings, design, sampling, data collection, analysis, reporting and other aspects of research conduct. The result of this is a detailed assessment tool, which is sensitive to the diversity of qualitative research but provides guidance for the appraisal of individual research studies.

8.16.1 Implications for government research, evaluation and policy making

No single piece of research, whether qualitative or quantitative, will point unconditionally towards a single policy action. But qualitative research, when properly conducted and credibly reported, will provide insight into the policy, how it operates, what makes it successful in

different circumstances, and how it sits and is experienced within the social worlds of those it affects. Users of research need to appraise critically the quality of qualitative research evidence and reporting. There are no standardised rules or procedures for doing this, but the user of research can expect a report to give them:

- a clear explanation of the research design and an account of its conduct, where the decisions made by the research team and their implications for the robustness of the data are discussed;
- a sense that the data have been critically and thoroughly analysed, with different data methods and sources used where appropriate. The reader should be left with a more nuanced and deeper understanding of the research topic, with new insights and conceptions of the topic;
- findings which are credible and plausible, rooted in the data, with the links between data and conclusions apparent;
- guidance as to the relevance the findings have beyond the study sample and setting, and enough description of the setting and findings for users to make their own judgements about whether the findings can be extrapolated to a different setting.

8.17 Recommended further reading

Bryman, R. (2001), *Social Research Methods*. Oxford: Oxford University Press.

Hammersley, M. and Atkinson, P. (1995), *Ethnography: Principles and Practice*. London: Routledge.

Mason, J. (2002), *Qualitative Researching*. London: Sage.

Patton, M. Q. (2002), *Qualitative Research and Evaluation Methods*, 3rd edition. Thousand Oaks: Sage.

Ritchie, J. and Lewis, J. (2003) (eds), *Qualitative Research Practice*. London: Sage.

Silverman, D. (2000), *Doing Qualitative Research: A Practical Handbook*. London: Sage.

8.18 References

Adler, M. & Ziglio, E. (1996) *Gazing Into The Oracle: The Delphi Method and Its Application to Social Policy and Public Health*. London, Jessica Kingsley Publishers.

Arber, S. (2002) Designing Samples. In Gilbert, N. (ed.) *Researching Social Life*, 2nd edition. London, Sage.

Atkinson, J. M. (1984) *Our Master's Voices: The Language and Body Language of Politics.* London, Routledge.

Audit Commission (2003) *Connecting with Users and Citizens*. London, Audit Commission.

Boaz, A. & Ashby, D. (2001) *Fit for purpose? Assessing research quality for evidence based policy and practice.* ESRC, Working Paper 11, London, UK Centre for Evidence Based Policy and Practice.

Bloor, M., Frankland, J., Thomas, M. & Robson, K. (2001) *Focus Groups in Social Research*. London, Sage.

Brannen, J. (1992) (ed.) *Mixing Methods: Qualitative and Quantitative Research*. Aldershot, Gower.

Brown, R. H. (1987) *Society as Text: Essays on Rhetoric, Reason and Reality*. Chicago, University of Chicago Press.

Bryman, R. (2001) *Social Research Methods*. Oxford, Oxford University Press.

Burgess, R. G. (1984) *In the Field: An Introduction to Field Research*. London, Allen & Unwin.

Cantrill, J. A., Sibbald, B. & Buetow, S. (1996) The Delphi and Nominal Group Techniques in Health Services Research, *The International Journal of Pharmacy Practice*, 4: 67–74.

Chambers, E. (2000) Applied Ethnography. In N.K. Denzin & S. Lincoln (eds.) *Handbook of Qualitative Research*, 2nd Edition. Thousand Oaks, Sage: 851-869.

Coote, A. & Lenaghan, J. (1997) *Citizens' Juries: Theory Into Practice*. London, Institute for Public Policy Research.

Critcher, C. & Gladstone, B. (1998) *Utilizing the Delphi Technique in Policy Discussion: A Case Study of a Privatized Utility in Britain, Public Administration, 76*: 431–449.

Davies, P. T. (2000) Contributions from Qualitative Research. In Davies, H. T. O., Nutley, S. M. & Smith, P. C. (eds) *What Works: Evidence-Based Policy and Practice in Public Services*. Bristol, The Policy Press.

Davies, S., Elizabeth, S., Hanley, B., New, B. & Sang, B. (1998) *Ordinary Wisdom: Reflections on an Experiment in Citizenship and Health*. London, Kings Fund.

Denzin, N. K. (1989) *The Research Act: A Theoretical Introduction to Sociological Methods*, 3rd edition. Englewood Cliffs, NJ, Prentice Hall.

Denzin, N. K. (1997), *Interpretive Ethnography: Ethnographic Practices for the 21st Century*. Thousand Oaks, CA, Sage.

Denzin, N. K. & Lincoln, Y. S. (2000) *Handbook of Qualitative Research*, 2nd edition. Thousand Oaks, CA, Sage.

Dingwall, R. (1997) Accounts, Interviews and Observations. In G. Miller & R. Dingwall (eds) *Context and Method in Qualitative Research*. London, Sage.

Fielding, N. & Fielding, J. (1986) *Linking Data*. London, Sage.

Fielding, N. & Lee, R. M. (1998) (eds) *Using Computers in Qualitative Research*. London, Sage.

Finch, H. & Lewis, J. (2003) Focus Groups. In J. Ritchie & J. Lewis (eds) *Qualitative Research Practice*. London, Sage.

Finch, J. & Mason, J. (1990) Decision Taking in the Fieldwork Process: Theoretical Sampling and Collaborative Working. In R. G. Burgess (ed.) *Studies in Qualitative Methodology* 2: 25–50.

Fishkin, J. (1995) *The Voice of the People*. Yale, Yale University Press.

Flick, U. (1992) Triangulation Revisited: Strategy of Validation or Alternative?, *Journal for the Theory of Social Behaviour*, 22(2): 175–97.

Fulop, N., Protopsaltis, G., Hutchings, A., King, A., Allen, P., Normand, C. & Walters, R. (2002) The Process and Impact of NHS Trust Mergers: A Qualitative Study and Management Cost Analysis, *British Medical Journal*: 325: 246 (3 August).

Geertz, C. (1993, first published 1973) *The Interpretation of Cultures: Selected Essays*. New York, Basic Books.

Gibson, T. (1998) *The Do-ers' Guide to Planning for Real*. Neighbourhood Initiatives Foundation.

Gilbert, N. (1993) *Researching Social Life*. London, Sage.

Glaser, B. G. & Strauss, A. L. (1967) *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago, Aldine de Gruyter.

Goffman, E. (1963) *Stigma: Notes on the Management of Spoiled Identity*. New York, Simon & Schuster.

Goffman, E. (1961) *Asylums. Essays on the Social Situation of Mental Patients and Other Inmates*. Garden City, NY, Doubleday Anchor.

Goffman, E. (1959) *The Presentation of Self in Everyday Life.* Garden City, NY, Doubleday Anchor.

Goffman, E. (1963) *Behavior in Public Places: Notes on the Social Organization of Gatherings.* Glencoe, Free Press.

Grewal, I., Joy, S., Lewis, J., Swales, K., Woodfield, K. with Bailey M. (2002) Disabled for Life? *Attitudes Towards, and Experiences of, Disability in Britain*. London, Corporate Document Services.

Hammersley, M. & Atkinson, P. (1995) *Ethnography: Principles and Practice*. London, Routledge.

Hayden, C., Boaz, A. & Taylor, F. (1999) *Attitudes and Aspirations of Older People: A Qualitative Study*. DSS Research Report No. 102, London, Department for Social Security.

Hodder, I. (2000) The Interpretation of Documents and Material Culture. In N. K. Denzin & Y. S. Lincoln, *Handbook of Qualitative Research*, 2nd edition. London, Sage: 703–716.

Honigmann, J. J. (1982) Sampling in Ethnographic Fieldwork. In R. G. Burgess (ed.) *Field Research: A Source Book and Field Manual*. London, Allen & Unwin.

Hughes, J. & Sharrock, W. (1997) *The Philosophy of Social Research*. London, Longman.

Krueger, R. A. & Casey, M. A. (2000) *Focus Groups: A Practical Guide for Applied Research*. Thousand Oaks and London, Sage.

LeCompte, M. D. & Preissle, J., with Tesch, R. (1993) *Ethnography and Qualitative Design in Educational Research*, 2nd edition. Chicago, Academic Press.

Lewis, J., Mitchell, L., Sanderson, T., O'Connor, W. & Clayden, M. (2000) *Lone Parents and Personal Advisers: Roles and Responsibilities*. London, Corporate Document Services.

Liebow, E. (1967) *Tally's Corner: A Study of Negro Streetcorner Men*. Boston: Little, Brown and Company.

Lincoln, Y. & Guba, E. (1985) *Naturalistic Inquiry*. Beverley Hills, CA, Sage.

Lincoln, Y. & Guba, E. (1986) Research, Evaluation and Policy Analysis: Heuristics and Disciplined Inquiry, *Policy Studies Review*, 5 (3): 546–565.

Lofland, J. (1971) *Analyzing Social Settings: A Guide to Qualitative Observation and Analysis*. Belmont, CA, Wadsworth.

Lowndes, V., Stoker, G., Pratchett, L., Wilson, D., Leach, S. & Wingfield, M. (1998) *Enhancing Public Participation in Local Government*. London, Department of the Environment, Transport and the Regions.

Mason, J. (2002) *Qualitative Researching*. London, Sage.

MacDonald, K. & Tipton, C. (2002) Using documents. In Gilbert, N. (ed.) *Researching Social Life*. London, Sage:187-200.

Morgan, D. L. (1997) *Focus Groups as Qualitative Research,* 2nd edition. Thousand Oaks and London, Sage.

Oxfam UK Poverty Programme (2003) *Have you been PA'd? Using Participatory Appraisal to Shape Local Services*. Glasgow, Oxfam GB.

Park, A., Jowell, R. & McPherson, S. (1999) *The Future of the National Health Service: Results from a Deliberative Poll*. London, Kings Fund.

Patton, M. Q. (2002) *Qualitative Research and Evaluation Methods*, 3rd edition. Thousand Oaks, Sage.

Philip, K., Shucksmith, J. & King, C. (2004) *Sharing a Laugh? A Qualitative Study of Mentoring Interventions with Young People*. York, Joseph Rowntree Foundation.

Popay, J. & Williams, G. (1998) Qualitative Research and Evidence-based Healthcare, *Journal of the Royal Society of Medicine*, 191(35): 32–37.

Ricoeur, P. (1996) *The Hermeneutics of Action*, edited by Richard Kearney. London, Sage.

Ritchie, J. & Lewis, J. (2003) Designing and Selecting Samples. In J. Ritchie & J. Lewis (eds.) *Qualitative Research Practice*. London, Sage.

Ritchie, J., Spencer, L. & O'Connor, W. (2003) Carrying out Qualitative Analysis. In J. Ritchie & J. Lewis (eds.) *Qualitative Research Practice*. London, Sage.

Robson, C., (2002) *Real World Research*, 2nd edition. Oxford, Blackwell.

Schaffner, C. (1997) (ed.) *Analysing Political Speeches*. Clevedon, Multilingual Matters.

Schuller, T., Brassett-Grundy, A., Green, A., Hammond, C. & Preston, J. (2002) *Learning, Continuity and Change in Adult Life, Wider Benefits of Learning Research*. Report No. 3. London, Institute of Education.

Scriven, M. (1991) *Evaluation Thesaurus,* 4th edition. Newbury Park CA, Sage.

Seale, C. (1999) *The Quality of Qualitative Research*. Oxford, Blackwell.

Seargeant, J. & Steele, J. (1998) *Consulting the Public: Guidelines and Good Practice*. London, Policy Studies Institute.

Shaw, I. (1999) *Qualitative Evaluation*. London, Sage.

Spencer, L., Ritchie, J., Lewis, J. & Dillon, L. (2004) *Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence*, 2nd edition. London, Cabinet Office.

Stewart, D. W. & Shamdasi, P. N. (1990) *Focus Groups: Theory and Practice*. London, Sage.

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

Strauss, A. L. & Corbin, J. (1998) *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*, 2nd edition. Thousand Oaks, CA, Sage.

Webb, B. & Webb, S. (1932) *Methods of Social Study*. London, Longmans Green.

Wengraf, T. (2001) *Qualitative Research Interviewing*. London, Sage.

White, C., Elam, G. & Lewis, J. (1999) *Citizens' Juries: An Appraisal of their Role*. London, Cabinet Office.

Whyte, W. F. (1943) *Street Comer Society*, 2nd edition. Chicago, University of Chicago Press.

Williams, D. D. (1986) *Naturalistic Evaluation: (Programme Evaluation Series 30)*. San Francisco, Jossey-Bass Wiley.

Wilson, J. (1990) *Politically Speaking: The Pragmatic Analysis of Political Language*. Oxford, Basil Blackwell.

Yin, R. K. (1994) *Case Study Research: Design and Methods*, 2nd edition. Beverley Hills, CA, Sage.

Yin, R. K. (1993) *Applications of Case Study Research*. Newbury Park, CA, Sage.