Part 1

Data Collection: An Introduction to Research Practices

1.1 Research and Data Collection

Doing research is much more than just gathering information or writing a description as a journalist would. Research consists in more intensive study, usually involving getting information that would surprise some audiences, and analysing the information carefully before writing up the results. The best research uses data in an original way or offers some new and exciting interpretation of existing data. Excellent research has to use evidence very carefully. Sophisticated data collection offers ways to plan and execute the best kinds of research. Many research outputs take a written form, but excellent research also results in soundbites that can be offered to practical audiences or in the media. These soundbites (punchy sentences) are only useful if the reader or listener can trace back to the original detailed research outputs, and perhaps even scan and scrutinise the data and data analysis themselves. The best researchers develop a strong reputation for systematic, logical and well-grounded research methods.

Some people argue that scientific research includes all the kinds of research that use data in sophisticated ways. But data are neither necessary nor sufficient for research to be scientific. Data are not sufficient because one also needs a carefully developed scientific argument. In this book, ways of developing good arguments are suggested and these rely in part on planning the whole research process before one starts to collect data. In some areas of research the phases of data collection and data generation are hard to distinguish because the data may already exist in newspapers or government publications, but one needs to be selective and propose that we use some of these resources. We then say that we are generating a dataset as a subset of the existing information. A good scientific researcher is likely to be able to generate or create datasets that are useful for scientific arguments. Data are also not necessary for all scientific arguments, because some arguments take the form of a normative or theoretical statement. Deductive arguments in particular may not require data at any stage. This book focuses more on other forms of scientific inference than deduction.

Research typically begins with the identification of a problem. After some general reading, one sets up a narrow research question which can easily be addressed during a constrained period of research. A literature review must be conducted and this review may include the close study of existing data and data analyses. The researcher then proceeds to gather fresh data or reanalyse and transform existing data. For most projects, a few weeks or months of doing more analysis usually follows. This book focuses more on the data-generation stages and less on the analysis stage, but the two are a little hard to separate because of the planning involved.

Systematisation is common in research. To systematically gather data might mean to run a parallel series of interviews on two sites, or to run several focus groups on the same theme. The choice of the research theme and the narrow research question is crucial. A few researchers in areas of sociology or philosophy may succeed merely by thinking about the issues and the works of previous thinkers. Even here, however, the works of earlier writers appear as a kind of data for the analyst. The vast majority of other researchers across the social and management sciences, medicine and health research, psychology and other topics have to collect and manage data as a crucial part of their research effort. Doing research may require the production of a project proposal to underpin a successful funding bid. Data collection may arise within the project proposal or may occur across a wider set of activities which we might call a programme. For example, one laboratory or institute may focus on how to utilise a longitudinal dataset or a set of cases arising from clinical meetings with patients. The research programme will then typically involve a series of smaller projects. Doctoral research often fits within wider research programmes. The degree of Ph.D. is awarded for scholarly research contributing new knowledge in a particular subject area. This degree requires between three and seven years of study. Other research projects take just weeks or months of work.

These brief notes on research do not do justice to the huge debate about what counts as scientific research. I have aimed here to introduce the various roles that data collection can play within the whole research process.

1.2 Findings

When a research project is written up and nearing completion there is often a moment of angst and concern about the main findings. Some of the stumbling blocks at this stage of a project can be over what to say, what nuances to place on different findings, who takes responsibility for these findings and how to integrate teamwork into an agreed document or presentation. The final stage needs to be foreseen during the data-collection stage so that when there are doubts, there is some recourse to the data or the data-analysis artefacts. Perhaps the data are a bit like the map that helps steering a course in a boat. The captain and crew decide where they want to go, then use the map to ensure they choose a reasonable and sensible way to reach the safety of harbour and complete their journey. Avoiding falsehoods, overcoming difficulties of comprehension, and translating between different dialects or lay idioms are all important ways that 'data' can help the researcher or research team avoid ending up like the *Titanic* – that is, at the bottom of the ocean.

The findings from a good study can usually be represented concisely on a single page in a diagram or other summary statement (as advised by Miles and Huberman, 1994). This advice given by Miles and Huberman was meant for qualitative researchers only, but it is good guidance for all kinds of social researchers. It helps to think of this aim as requiring conciseness, focus and a certain narrowness of the main topic of the research. Most researchers base their 'findings' closely on their research question (see Wisker, 2008: ch. 5). Some, however, revise the research question as they go along. These people tend to get into difficulty when writing up because it may become unclear what exactly they are focused on. Therefore, in writing up your findings a good guidance is first to answer the original research question and then make additional comments about exploratory aspects of the findings and new proposals for future research.

It is often easier for a lone writer to achieve a good write-up or presentation because they do not have to be monitored or influenced by others. On the other hand, the lone author runs a risk of making arguments that others will find ridiculous or unsubstantiated. It is always a good idea when developing a draft to ask at least three people to read it early on. For those working in teams, individuals can write drafts and sections and pass them around. Guidelines for co-authoring can set out the roles team members may play (British Sociological Association (BSA), 2002). For example, one person might be a ghost writer and not want attribution, while another who collects data may want to be a named co-author. In general, the BSA tends to discourage ghost writing and suggests instead that the lead author may have a number of named co-authors, including the paid writer, who then get to claim joint authorship and take a fuller role in finalising the report. The BSA also encourages data enumerators and interviewers to become named authors. The guidelines argue that to be a named co-author of findings, each person needs to be aware of the whole paper's arguments and to have gone over the whole text in detail at a late stage to make comments, insertions and corrections. As long as this has happened, some co-authors can have a relatively minor role in writing up but may have had important roles during data collection.

Some findings will probably 'emerge' from a study without being expected or predicted in advance. The 'expected' findings might follow the usual pattern of normal science. 'Normal science' is a special phrase referring to the use of standard, pre-planned methods on a routine study topic to get results that to some extent (in broad outline) could have been predicted from the start. Kuhn (1970) saw normal science as rather conventional and pointed out that the very best science goes beyond normal science. Anomalies and situations that are new, unique or difficult to explain cause scientists to seek new, creative, innovative explanations or interpretations. Now the same dataset could be the basis of a new set of findings! This exciting vista, set out by Kuhn as paradigm change but also studied by many other authors since Kuhn's time (Fuller, 2003), offers researchers a wide range of ways to deviate from normal science.

Although there are connections (and rootedness) between the data collected and the findings, there is not a single mapping from one to the other. If we stay with our sailing analogy, there are many ways to reach the other side of the ocean. Social science archives (where data are held for future users) offer opportunities for reanalysing older data, comparing it with new data and perhaps applying new interpretive strategies. Thus there is not a single set of findings from one study. Tolerance, pluralism and even-handedness are needed when we realise that different researchers might develop different findings from the same dataset.

According to Kuhn (1970), the hypothesis-testing tradition led to a great pretence of attempts at falsification when in fact the underlying basic theoretical premises were never questioned. As a result, some scientists now avoid the hypothesis-testing methodology. I offer several approaches to this issue in this book. For example, you could be a qualitative researcher developing new hypotheses. You would, of course, test them at the same time and offer sensible claims. But no quantitative research would be involved. If you were a quantitative survey researcher you might test a wide range of hypotheses and tell the reader what you have found out. A mixed-methods researcher has to weave a pathway delicately around these options. Some researchers now do a little of both. In rough terms we might call the first 'exploratory' findings and the second 'tested' findings or simply 'hypothesis-testing'. In order for it to make sense to do both, one needs to leave the tradition described by Popper in which falsification played a central role. One should take responsibility for choosing theories and decide on which set of basic assumptions to work with. Kuhn argued correctly that falsification was not value–neutral in general (Fuller, 2003, 2005). Researchers today, such as the textbook authors on research design, see mixed methods as highly feasible and desirable (De Vaus, 2001; Teddlie and Tashakkori, 2009; Creswell and Plano Clark, 2007; Creswell, 2003, 2009). Advice has tended to move away from the traditional separation between a value-neutral survey method and a value-saturated qualitative method.

I wonder whether the word 'findings' tends to suggest a consensual set of conclusions and so denies the possibility of contesting the results? In recent years it has become clear that many of the basic words and phrases used in social science are essentially contestable. Examples such as human rights, civil society, rational choice and market optimum are obviously controversial for those who do (or do not) adhere to the underlying values that they imply. Social science is not generally considered to be value-neutral any more. Specific concrete sentences may be factual but an overall argument usually has values (whether explicit and/or implicit), a purpose, underlying assumptions and a persuasive undertone (Olsen, 2009).

The most exciting writing certainly has a purpose. This chapter on findings is meant to excite you about the prospect of doing research, even knowing in advance (with trepidation) that the findings are going to be contestable! Having this clear purpose, I can write in a well-informed, focused and informative way: that is what good research is like too. Researchers use evidence as part of their arguments but in a way that other writing does not. So there are necessary connections between the data-collection plan and the goal or horizon of the kinds of findings that are expected from your study.

To summarise: research findings may be controversial but need to fit into an argument that is clearly stated, well grounded with evidence and suitable for further discussion and development by others. Research data can be put into a data archive to enable early findings to be reassessed later. Using tolerance, researcher teams can develop agreed findings even when they are not wholly unanimous about some aspects of policy or values. Using pluralism (which will be discussed in more detail later on), researchers can explore the usefulness and validity of competing theories in an environment that accepts that more than one theory may have a 'purchase' on events or a telling insight to offer. Hypothesis testing and falsification are not the bread and butter of social science even if they are, at times, very important.

1.3 Data

Data are disembodied information. Data are not the same as knowledge. My favourite type of data is interview transcripts. These are rough and raw – often embarrassingly so – but they reveal a lot about a scene to the close reader. The data type I use most often is survey data. Examples of these two data types, shown in Table 1 and Box 1, illustrate how top-quality researchers in the UK frame these two types of data.

Table 1 BHPS data on ID, age, sex, marital status, and flexitime – six-row sample

pid	age	female	married	flextl
167020423	25	yes	yes	0
958518444	51	yes	yes	0
168083472	45	yes	yes	0
520740483	44	yes	yes	1
971938955	45	yes	yes	0
115014131	49	yes	no	0

Note: The data are anonymised here.

Box 1 Extract from a Transcript of a Three-way Interview

Topic: Television industry contractual terms

Length: 18 pages single spaced

Interview ITV Company 15 March 2000

Interviewees:

Person 1: Researcher in the TV industry, female aged 22. Single. Recently graduated from Cambridge with good degree in politics. Short term contract.

Person 2: Post-production/video tape library, male aged 43. Married with one child. Permanent staff in the TV industry.

'Me' is the interviewer Valerie Antcliff.

Me: The first thing I'd like to ask you is how secure you feel in your current position?

Person 1: Not very at all! I graduated last year so this is my first sort of proper job and to begin with I was clearly – oh well this is what happens, you're on three month contract, I mean I've just been told that my contract ends at the end of April and my next contract sort of goes up until the end of June and that's it, so essentially from June I'll be unemployed. Now that's because the series I work on will finish in June, but other things will come up that I can possibly go on to that will last over the summer, but there isn't that guarantee and actually it is rather an odd feeling, yes technically I could be unemployed in June. The way I've sort of looked at it is, well I do actually enjoy the job I do but if something else comes up I'm not tied to it because I've not signed a year-long contract like a lot of my friends who got jobs when they graduated have. They sign these great big, long contracts to sort of be there for ever and ever and me, I'm kind of in the reverse of that. But it is slightly unnerving because there is that sense that you think well if they think I'm totally terrible, well they will just get rid of me overnight.

Me: So do you think it could work to your advantage? If a better job comes up....

Person 1: It works to my advantage because I come into the job knowing I'm not going to be there for very long...

These two examples are chosen to illustrate the two extremes of 'hard' and 'soft' data. People usually mean by hard data the highly structured questionnaire data types. The hardness comes from answers being placed rigorously into categories, as seen in Table 1. These categories are set up to have the same meaning across the whole sample. In the case of the British Household Panel Survey (BHPS) this would be across the whole United Kingdom (Taylor, 2001). Little adaptation is possible within the table, called a 'matrix of data', to allow for possible variations in Wales or Scotland compared with England or Northern Ireland. Thus a kind of universalism of meaning within the population is assumed in so-called 'hard data'.

The 'soft data' of the interview is framed sequentially by a delicate balancing act of the three actors – the female interviewer Valerie Antcliff, the female respondent and the male respondent. All interviews vary in their content, and words are not assumed to have any specific (stipulated) meaning. Even a semi-structured interview like this one has a softness to it in the mapping between words and meanings. Instead of universalism, there are assumptions like concrete local specificity, subjective variations of meaning and exploration of tacit meanings in the construction of the interview. Interviews have flexible sequence so that any idea or claim can be explored. In this example we see that the idea of job security is being explored by the respondents.

An interview is a mutual construction. By contrast, the content of a highly structured questionnaire – and its flow – are dominated by the people who constructed it.

In the interview, any attempt to change the subject, dominate the flow of topics or push further into a given subject is evidently going to be 'led' by somebody. It is a very personal, messy, interactive business. In a survey the 'leading' is done institutionally at the questionnaire preparation stage. This 'leading' is usually managed through careful piloting (i.e. interviews) around blocks of draft questions. Once the questionnaire is printed up, the questions and the flow leading away from certain topics are fixed for all respondents. To a respondent, the leading is invisible. It has hardened into printed instructions, which may seem rather authoritarian. For the questionnaire 'enumerator' or face-to-face interviewer there is very limited room for manoeuvre.

The survey method is also not 'soft' in the way it is delivered. Surveys are often delivered by hand and then collected by post, or even conducted on the internet. But the best survey data, like the BHPS, are collected by hand during a face-to-face interview. This interview is so institutionalised that some of it is now captured immediately on a laptop computer. The face-to-face method enables the questioner to help the respondent navigate the flow of the survey. The questioner can also clarify any ambiguities that may arise. The basic idea of a survey, though, is that the viewpoint of the questioner (enumerator) cannot colour the way the questions are phrased. This supposed absence of the enumerator's subjectivity helps ensure national or even international homogeneity of the data that are recorded. The homogeneity of meaning is only valid, of course, if it is both feasible and realistic to use a single question to tap into an area of experience right across the whole space being sampled. The survey method depends heavily on this homogeneity assumption. By contrast, interview methods are robust to local variations in meaning. Interviews can also handle the situation where people hold contradictory views at a single moment. In a survey, by contrast, as everyone knows, the respondent is supposed to put up a consistent self-image. The data simply records that image.

These two data types – surveys and interview transcripts – are sometimes confused because we often use an interview to fill out a questionnaire. The usual way to distinguish them is to say that structured interviews are used with questionnaire surveys, but semi-structured and unstructured interviews are used to generate interview transcripts. For an unstructured interview you might not even produce a transcript, preferring instead to follow a topic through a series of visits to one or more respondents.

There are many different data file types that correspond to the different forms of data. Here are some of them:

- documents, with file extensions .doc, .rtf or .txt, for interview transcripts;
- images, with .jpg, .tif or .wmf, for photos;
- spreadsheet data, with .sav, .dat or .xls, for survey data;
- sound and video, with .mp3 or .mpg, for interview sound files.

I will take a moment to briefly describe each of these types of data. Note, however, that there are many other types which are not computerised! Examples include handwritten maps, focus-group tape recordings, art drawn by children, interim action research memoranda, and field notebooks. As this book progresses, most of these data types will be discussed.

The .rtf format is a substitute for a Microsoft Word document. RTF stands for 'rich text format'. This format is transferable between computer packages, and can retain some computer graphic images, such as logos, even when moving away from the original software. There are some formats the .rtf file cannot be moved into, such as Adobe Acrobat .pdf format, without buying the proprietary software that corresponds to that format. Adobe Acrobat .pdf format files are nicely encapsulated to present the graphics, page margins, font and footnotes consistently, no matter what computer the file is opened on. The .pdf file is highly portable because the user does not need to have Microsoft Word software. Indeed, the user typically cannot edit the .pdf file, and is a passive reader rather than being a co-writer on the document. As a result, the .pdf format is very common but is not usually a desirable 'data' format and does not appear in my list. Instead, the older and more basic standard '.txt' format appears. The .txt file is simply a text file and cannot hold graphic images or logos. Microsoft Word can create a text file but it will drop complex layouts such as page numbers or footnotes. From a .pdf source file, using cut and paste, one can create a .txt file that contains usable data. The 'usable' data are easily edited, moved, translated, annotated, and so on: it is a writeable as well as readable file. The .doc format is usually fine for such text files. For open source software users, a variety of other file types can be used too.

The images taken by digital cameras are mixed in with logos and other computerised pictures and graphics in the filetype .jpg. Many other file types exist for holding images, including bitmaps (.bmp) and tagged image format files (.tif).

Social scientists often learn to use a spreadsheet during their basic training. The spreadsheet is a table with numbered columns labelled A to Z and AA, AB, AC, and so on. Potentially, several hundred columns and many rows can be filled with detailed data. The data can be words or numbers. These spreadsheets take a special format in social science, where the numeric data can be mapped into a coding scheme to give each value a word or phrase as its value label. For example, 'yes' would be 1 and 'no' would be 0. In the example in Table 1, the variable flext1 (‘flexitime work arrangement in paid job’) is coded thus. The coding scheme for gender (female = 1 and male = 0) and the coding scheme for marriage have been entered into the computer, giving words in the table but with numbers stored in the computer spreadsheet. In this case I used STATA software (see Hamilton, 2004; STATA, 2003). Because the coding scheme for FLEXT1 is not yet entered into the machine, the numeric values show up in the table. I can declare the value labels, and the table will then appear as shown in Table 2.

Table 2 BHPS data after further coding

pid	age	female	married	flextl
167020423	25	yes	yes	no
958518444	51	yes	yes	no
168083472	45	yes	yes	no
520740483	44	yes	yes	yes
971938955	45	yes	yes	no
115014131	49	yes	no	no

Note: The last column now shows the value labels rather than the raw code values.

A dictionary summarising the coding can be produced from the special software such as SPSS (Field, 2009). One of the simple tasks this package does is to produce a list of variable names – which are not the same as value labels – as shown in Table 3.

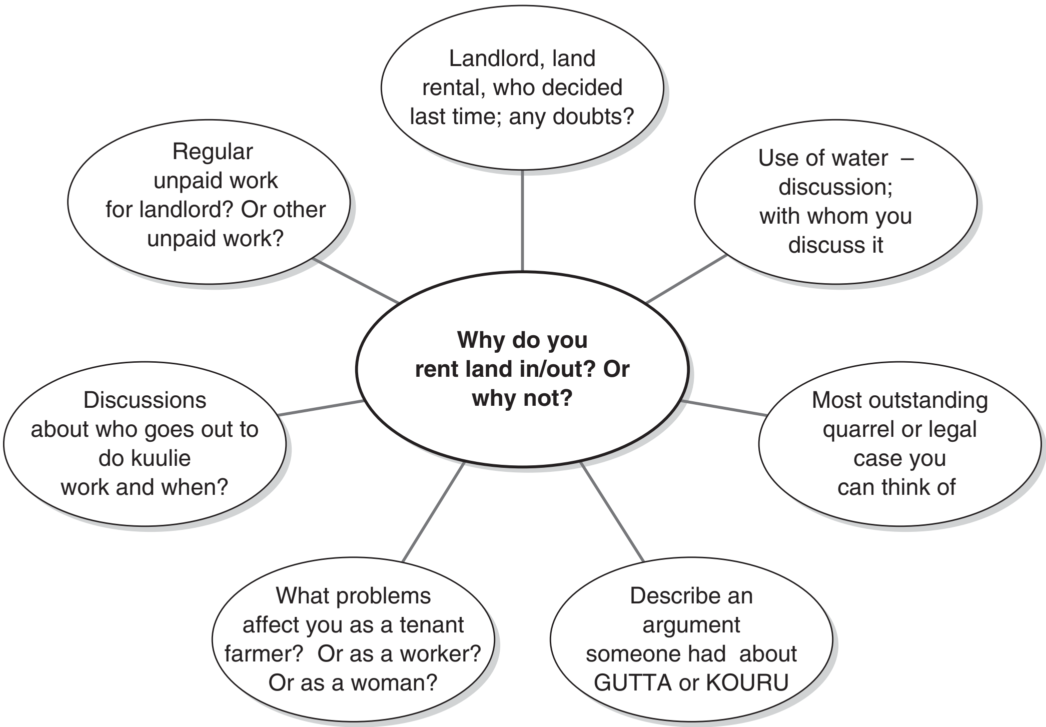
Table 3 A data dictionary segment

Variable Name	Variable Label
pid	Person identifier
age	Age at Date of Interview
female	Female
married	Married/Living as Couple/Partnership
flext1	Work Arrangement: Flexitime

Sound and video files can be held temporarily on a minidisc player or MP3 player. Dictation machines can be used to record sound. The .mp3 or other sound files can be uploaded to a computer afterwards. Another alternative is to put a microphone into the input socket of a computer and make a recording directly onto the personal computer. This will generally appear as an .mp3 format file. Video files take many formats and researchers using a digital camera may decide to make short shots of video using .mpg format to capture a glimpse of respondents – with their prior permission, of course. Observation studies can use the .mp3, .mp4 or .mpg format to keep a detailed video record, including sound, of various activities, again with prior permission. Compact discs (CDs) generally use a different file format such as the .cda format. The material on CDs can be ‘ripped’, i.e. uploaded with a new format, to make an .mp3 file. If the source data are speeches recorded on CD, then a simple computer operation can rapidly create .mp3 files from the original files. Similarly, sound on phonograph records or cassette tapes can be converted.

During a research project computer data files are generally complemented by handwritten 'field notebooks' that are gradually filled by each researcher during a project. Field notebooks pile up over the years and form a crucial backup copy of raw data, contact addresses and the researcher's initial thoughts. I usually start off a field notebook with my name and address on the front so that if it gets lost, it – and the crucial research notes, phone numbers and ideas I've jotted down as I go – can be returned to me. Then I put the 'research question' – the one that identifies what is really special and unique about this study. Next, I start making lists of contacts and what they do and where, when and how to reach them. When I visit an 'informant', I open the book and say that this is where my notes go. If they want to remain anonymous I use a pseudonym for the notes pages. I keep a list that matches pseudonyms with real names in a separate, secret place. The field notebook gradually fills up with notes and reflections. The interview plan is taped into the notebook, interviews that are taped do not go into the notebook, but I will make a note of the cassette number or MP3 file name and who I have spoken to, when and so on. The notebook can then be used as the question source and it is no longer necessary to take notes once the data are recorded as sound.

A sample of an interview plan is shown in Figure 1. This plan is very concise and would be put into the notebook and handed out to respondents. Data from


Landlord, land rental, who decided last time; any doubts?
Regular unpaid work for landlord? Or other unpaid work?
Use of water – discussion; with whom you discuss it
Why do you rent land in/out? Or why not?
Discussions about who goes out to do kuulie work and when?
Most outstanding quarrel or legal case you can think of
What problems affect you as a tenant farmer? Or as a worker? Or as a woman?
Describe an argument someone had about GUTTA or KOURU

Key: Gutta = renting land on a cash rent basis;
Kouru = renting land on a sharecropping basis.

Figure 1 Plan of a semi-structured interview

this particular plan were used for writing several short articles. There were four interviewers: three native Telugu speakers and me. We did not try to make all the interviews the same, but we tried to cover all the topics shown in Figure 1 and Box 2 in each interview. I will discuss this exemplar more later.

Box 2 Interview Plan Used in Rural India

INTERVIEW PLAN EXTRACTS

Would you agree that we tape your interview so that Wendy can improve her knowledge of Telugu and also write about the situation in the villages? _______Thank you.

Were there any doubts about decisions that were made last year about the renting of land?

How and when was the crop choice made last time, for putting crops on the tenanted land? This can refer either to last year *kharif* or to the current *rabi* season.

Basically why do you rent land?

(or why do you not want to, if you have doubts about renting it?)

[8 questions omitted]

Think of a situation when someone wanted to do kuulie [casual paid] work, and there was a disagreement about it. Tell me about that.

Think of a situation where it is routine to do kuulie work. Tell us who decides about that.

[7 questions omitted]

Etc. prompting till some disagreements are described, and some household-level agreements are described.

[3 possible supplementary questions omitted] End.

Note: The interview plan is shortened for the sake of brevity. *Kharif* and *rabi* refer to the winter and spring planting seasons, respectively.

In discussing ‘data’ at a more abstract level I will now consider impartiality and validity. Impartiality is important because most people think that hard data create an impartial and unbiased source of information. This is not generally true. Being disembodied records, the ‘data’ do have three qualities that make them less personal than the usual face-to-face interview:

1. Reliability. No matter which enumerator did the questioning the answers are still harmonised around the same core meanings.
2. Validity. During piloting there were careful checks that the meanings are consistent or homogeneous among different subgroups of the population, e.g. young and old, north and south, and different language groups.
3. Authenticity. There are checkable, disembodied historical artefacts recording what 'was actually said':
   (a) The advantage of the survey method is that we know exactly what has been recorded. *X* per cent of respondents said *This* in response to *That* prompt. 100 – *X* per cent did not say *This*.
   (b) The disadvantage of the method is that we cannot be sure what they meant to convey by giving or choosing that response. Close-ended multi-answer questions are the worst culprits here.

Data as historical artefacts have great advantages for the study of change over time. But in a revisit study or 'longitudinal study' the questions would have to be framed the same way the second time around for the comparison of answers over time to be valid. Validity and verification are important areas for survey research.

Reliability has a specific meaning in the context of social research. It means that a study's findings would be the same even if it had been conducted by a different person. There are two levels of reliability: the reliability of data and the reliability of the findings of the survey.

When considering the reliability of data, the survey method helps to ensure harmonisation and homogeneity. The delivery methods also help. Having literate respondents complete their own questionnaire is a useful method, termed 'self-completion'. If interviewers are used, they should be non-interventionist and should not vary the wording of questions. They should not introduce their own opinions between questions or at stages of the questionnaire. If they do, this will reduce data reliability. The data obtained from each delivery of the questionnaire should not vary from one enumerator to another.

The issue of the reliability of the findings of the survey, on the other hand, is much more problematic. Given a single survey data table, each separate author might write divergent themes and findings from that table. Their differences might arise from theoretical differences, cultural assumptions or selective use of different parts of the data. Beginners often think that findings or results should be reliable. I would advise that this position reflects a misunderstanding of social science. The results might be reliable between different interpreters, but this is not necessarily a good thing. When a new interpretation arises, it can be invigorating, democratic, interesting, innovative, helpful and creative! So I do not idealise 'reliability of findings' in the same way that most survey researchers aim for 'reliability of data'.

Validation should not be confused with reliability. Two forms of validity are often referred to in survey research. Firstly, internal validity occurs where measures conform to the stipulated meanings that the researchers intend to associate with the words used. Internal validity can be developed through a convincing argument about how the questions used relate to the research question and the topic of research in general. Internal validity also requires consistency among a research team's writings. The meaning of key terms must not be changed in mid-stream.

Internal validity is contrasted with external validity. Here the data are supposed to be constructed in such a way as to have consistent meanings both for the researchers and for the respondents. It is also important that the intended audience of the research understands the central concepts in consistent ways. External validity is hard to achieve in a world of contested concepts such as 'social capital' and 'cultural values'. Some social science disciplines like psychology and management put higher value on external validity than other disciplines do. Sociology and politics are particularly willing to allow for multiple interpretations of the same data. If multiple interpretations are possible, and different standpoints of interpretation have their own way of constructing their internal validity, then the phrase 'external validity' will not make much sense. In these disciplines, and for qualitative researchers more generally, there is sometimes impatience with the sweeping generalisations made by survey researchers. Nevertheless even the qualitative researchers have concepts of validity. These are described elsewhere in this book.

Having looked briefly at reliability and validity, we come back to the basic core concept of 'data' in social research. Data are artefacts. Datasets are like mini-museums. Carefully constructed, managed and displayed, data can be a rich resource both for knowledge and imagination.

Data archiving helps retain these data artefacts for future generations and for use by outsiders. Many countries run a national data archive. In the UK the institutions include the Economic and Social Data Service (www.esds.ac.uk) and the UK Data Archive (www.data-archive.ac.uk). The US Census is held on a data archive (www.archives.gov). International organisations such as the International Labour Office and other parts of the United Nations manage data release as a regular part of their operations. The ILO's data can be found online at www.ilo.org. These archives are carefully managed to retain verifiable links between the survey documentation, the validation or piloting activities and the sampling methods used and the datasets themselves. Data archives for survey data help to ensure that we can retain the anonymity of respondents even in the face of concerted attempts to link up datasets to discover people's addresses or identities.

On the qualitative side, too, new data archives have been created to enable the retention and reuse of interview data and other qualitative data types. Both qualitative and quantitative data archives offer some resources free to academic users. Commercial users and those wanting data on hard media and other services are typically charged fees at around the cost of production. These days costs can be minimised by having data transferred through the internet. A common method, known as file transfer protocol (FTP), allows the archive to release the file to the user for a limited period, for example one week, using a password sent to the user via email. The user simply downloads the file from the Internet. The user is restricted in their freedom to copy or pass on the data to others. A signature is usually required to ensure they agree to all the restrictions on their use of the data. Using FTP means the transfer of the data can be extremely quick but might be preceded by a period of a week or so during which emails and signed documents are passed around between user and provider.

So far I have discussed the management and validity of some data types. There are no hard-and-fast rules about the validity of data, so ‘beauty is in the eyes of the beholder’. Each researcher will have to develop their own viewpoint about validity and the knowledge that can be gained from research using data. Within a team, there may even be disagreements about knowledge. Ideally each researcher will have well-developed views about it.

1.4 Causes

For some researchers, planning a data-gathering project revolves mainly around notions about how to explain certain outcomes. An explanation involves cause and effect relations. Both survey data and interview data can help us learn about causes and their effects. In both quantitative and qualitative social research there are also times when we do not want to study things causally: we may want to just explore a phenomenon or look at all aspects of it. But when we do want to look at the causes of things, it gets very complicated very quickly.

In this chapter I will define a cause and give some examples of causes and causal factors. Then I will illustrate the data that we collect about such situations and raise some implications of the 'epistemic fallacy' so that you will not get caught out resting an argument too heavily upon a spurious correlation.

A way to define causes is to recognise that social life is changing over time, and that all new events have their origins in the past. We call something causal if it is one of the factors that contributed in the past or present to an event in the present. Causal conditions include structural factors such as the social class you were born into, and, perhaps, are still in; government policy and institutions; and the kind of city you live in. Furthermore, in any chain of events there are direct and indirect causal factors.

There are growth spurts, and the food for growth is causal. There are newly emerging institutions, and the institutional history is part of the causality of the new ones. There are unique and new events, and the creativity of the inventor may be causal there. There are obstacles to certain outcomes and whatever removes the obstacle is a cause of the outcome. Finally, there are enabling factors. These are causes but often only work in the presence of other factors. So you can see there are many types of causes. Some people like to think of deep causes, indirect causes, direct and proximate causes. Of these, the proximate cause is the one nearest to the outcome in time or space. But actually it is hard to distinguish these types of causes because some causes are simply background factors which coexist with – and thus pre-date and post-date – specific causal events. It may be hard to say which is most important. Discerning causes has a subjective element to it.

For me, a cause is mainly interesting if it is not both necessary and sufficient for the outcome. For some empiricists, the opposite is true: for them a cause is only interesting if it is proven to be a cause by being both necessary and sufficient for the outcome. I look at this issue separately and in more depth in the chapter on **causal mechanisms** (6.5) in Part 6. If something is always necessary and sufficient for an outcome, then it is just part of that outcome's pattern: 'letting the lions into the ring' is causal for 'having a lion show in the circus'. It does not really explain why one circus has a lion show and another does not. Taking divorce as an example, marriage is a necessary condition for divorce but is not really an enabling factor since marriage simply lies in the background of all marriages, whether they ended in divorce or not. The type of marriage, date of marriage, mode of marriage, and so on are much more interesting factors in explaining divorces. Divorce papers are not a very interesting cause of divorce either. These are the papers that one partner receives, usually by post, to inform them of the divorce proceedings occurring. They are of course a cause, and both necessary and sufficient to 'cause' divorce, yet the divorce papers are inherently part of divorce. They are not a contingent factor at all. I should point out that this divorce example is based on Western Judeo-Christian systems where divorce goes through a court, not on Islamic systems where the paperwork is not as important.

In summary, the most interesting causes are contingent causes, not inherent background factors. Because causes are contingent and outcomes may occur either accidentally, unpredictably or only in certain patterns or mixtures of causes, discerning causes from empirical data is quite tricky. Even when causes A, B and C are real, they do not work in tandem all the time to cause outcome Y. Therefore evidence about the co-incidence of A and Y may be misleading about whether A causes Y. Even if A really causes Y, it may not co-occur with Y in all situations. If there is a partial association of A and Y, does that mean we refute that A can cause Y? I do not think so. But on the other hand, data about the association of A and Y do not 'prove' that A causes Y either. Proofs of causality are hard to develop.

To be really definite about this, let us look at three examples of causality. Firstly, breast cancer does not always cause death. But it can cause death. So I say: 'Cancer tends to cause death.' This places the causality as a *tendency* of the cancer, not as a deterministic statement or as a fact about all cancers. Intervening factors such as discovery, treatment, food intakes, prior health, and co-occurrence of another cancer are all relevant causal factors. These make data collection a complex task for studies about death from breast cancer. Studies of epidemiology – the causality of illness – stress that if we are to successfully and thoroughly study the causality we need to have survivors of breast cancer in our study as well as those who died (Kreiger, 1994). Studying the causality of death from breast cancer can lead to a strong focus on healthy lifestyles, because survivors may tend to share one common trait, such as healthy lifestyles or a healthy attitude to their body. As examples of this, women with healthy lifestyles survive chemotherapy better than those with prior health problems. Those women who check their breasts frequently are more likely to detect a small cancer before it becomes life-threatening.

Secondly, in social work, it is often observed that poor school outcomes are associated with broken families and economic deprivation. But what is causal in those children's backgrounds? The 'nexus' of their conditioning is complex and has developed over many years. Their interaction with teachers is affected by the teachers' experience of local schools. Maybe the teachers are the cause of poor school outcomes for such children. Maybe teachers cause some outcomes, and children cause some outcomes, but the poor performance outcomes are the same ones. It becomes hard to isolate and measure any single outcome.

Some research papers argue that divorce is bad for children's schooling, while others say that marriage is positively good for schooling. It is important to see that these are different arguments, not merely symmetrical opposites. On the other hand, there is also evidence that children with richer or service sector parents do well even if there is only one parent. (One detailed illustration is provided by Fuller et al., 2002.) The social worker wants to know whether interventions in a child's life can improve school outcomes. But the interventions are themselves unique, personal, interpersonal, hard to record (in detail), and often unrecorded. Scientific survey research may not capture the real effects of transformative meetings of children with social workers and may falsely draw attention to spurious causes, such as the parents getting a good job, which actually arise (also) as outcomes of the social workers' interventions. Thus there is complexity.

Speculation about all these causes is actually a very good way to plan and proceed with widening the data-collection process. That way, no important factor will be left out. Qualitative researchers tend to argue that following single cases over time is an excellent way to understand what factors are transformative. For them, the causality that matters is what transforms a child 'inside', and what makes a household begin to work more functionally (i.e. be less dysfunctional). These things are unique. They also want to understand the dynamics of classrooms and do not think this can be simplified down to 'class size' or the 'curriculum and testing regime'.

Thirdly, in economics we get notions of causality at other levels; that is, beyond the individual. Here you may find that the bankruptcy of a single bank causes a wave of fears of recession and that these fears then cause a strong recession, and then that recession causes unemployment and lower incomes. The causality cannot work the other way, says the economist. But the argument about causality of recession is still contested. The same recession could be attributed to any number of factors: monetary policy, industrial policy, global competition or a protective trade policy on imports and exports. In economics, as in other areas of life, there are so many potential causes that explanatory arguments are highly contested. The role of markets, the emergence of new characteristics in markets and, in general, large-scale 'aggregate' or social properties of markets are offered as alternatives to individualistic explanation (Lawson, 1997). Some economists ultimately commit themselves to reducing everything to individual action by tracing large-scale phenomena backwards until some individual action is found. For example, a poor lending decision by a banker might have indirectly caused the bank to go bankrupt, or a poor risk-management policy by a bank policy-maker might have contributed to the wave of fear, and so on. Again, complexity rears its head, because we have so many different actors on different levels. In economics, as in health and social work, there are multiple stages rather than just a single drama with a few actors.

These three examples show that many causes contribute to outcomes, and that focusing on an outcome and its causes leads us toward studying the history of the present day. We also see that factors exist at many levels which all contribute towards the configurations – or situations – that we are trying to explain (Lawson, 1989).

Ultimately a 'cause' is best thought of as a mechanism which has the capacity to generate some outcome. This capacity is a tendency, but is not always activated. Other causal mechanisms, such as social structures and institutional factors, also intervene. Therefore, the context is part of how a causal mechanism actually generates a particular outcome. For further reading about causality and its measurement I recommend Pawson and Tilley (1997) and Pawson (1998). The latter is quite critical of survey measurement. Summary arguments can be found in Byrne (2002). Nevertheless the consensus is that causality exists.

In the British Household Panel Survey, numerous incidents in a person's life are recorded year after year through a series of annual house visits. The first visit includes a short life history of the person's employment experience, the second visit consolidates this history, and then annual visits create a panel dataset in the standard survey-data format. The BHPS is unusual in having different segments – some done as face-to-face interviews, some using computer-aided interviewing and some using self-completion questionnaires in the home. The rows of data – one row per case – are computerised and made anonymous. We can then trace through the sequences of events from 1991 to 2007 in the lives of any of thousands of respondents. If you are at school or university, a teacher or researcher, see www.esds.ac.uk for free access. Commercial users have to pay a fee.

This is how we can organise the BHPS data assuming variables can act as traces of real causes. The BHPS columns are the variables, and I can list several types of causal variable from BHPS (Table 4). The causality of an outcome such as divorce or poor school performance might be constructed by looking at all the other explanatory factors shown. The outcome might be caused not just by personal but also institutional and social factors as listed.

Table 4 Variable types drawn from the BHPS

Structural factors	**Institutional factors**	**Personal factors**
• social class of father • social class of respondent • service class job • household income • age group • marital status • ethnic group • region • rural/urban • home owner or tenant	• whether employer has a union • hours worked per week • whether the hours are flexitime or not • whether job has a pension • whether attends church • whether a member of a voluntary organisation • whether self-employed	• whether respondent is a member of a union • whether person wants to work more hours per week than they currently work • gender • response to an attitude question about managing early-years child care • division of domestic work

So if we put to an interviewee the question 'How strongly do you agree with the statement that a young child suffers if its mother goes out to work?' we need to prepare for a much wider analysis. We cannot just pick any causes we like out of a hat. We need to gather data on structural, institutional (e.g. membership) and event-based causes of different attitudes in answering this question (see Crompton et al., 2005; Crompton and Harris, 1998).

The epistemic fallacy is the tendency to confuse the data with the reality. This fallacy often starts by a statistical method which focuses on given data and does not collect primary data. Being restricted in the variables available, the researcher tends to look only to these to get answers to an explanatory question. The fallacy arises when cause is attributed to what is present (e.g. ethnicity), rather than to the deeper unrecorded factors (e.g. discrimination). In the 'child suffering if mother goes out to work' question, the recorded cause of a strong belief that children suffer might be found in the variable 'age', when the real cause is much deeper, i.e. beliefs about housewifery that were taught to the older generation when they were young. Those who have new training or ideas will tend to disagree with the statement, and those who retain their beliefs over a long period may tend to agree with the statement. To attribute cause merely to 'age' is to misspecify the cause. Bhaskar (1989) is attributed with calling this the epistemic fallacy. His contribution was to show that retroduction can offset the dangers of the fallacy of using only pre-given data. Following loosely Bhaskar's line of thinking, we can advise that retroduction plays a part in scientific research using a typical research design or protocol (shown in Box 3).

Box 3 A Protocol for Hypothesis Testing within a Petroductive Methodology

- State the theory or theories.
- List the hypothesis or hypotheses.
- Collect and analyse the data.
- Test the hypotheses.
- Do more data collection to explore the situation.
- Reflect for a moment.
- Frame new hypotheses.
- Revisit existing theories and draw conclusions.
- Start the next project.

This protocol for hypothesis testing within a retroductive methodology accurately describes what many researchers do, whether using survey data or not. It encourages all researchers to acknowledge their search for a broad expertise and for qualitative and theoretical knowledge. Getting more knowledge (during the middle of a project) is obviously desirable as the learning process proceeds. Even without individualised survey data, we can study history and documents to understand the real causes behind a pattern. Unpicking the epistemic fallacy was a major contribution to the improvement of research designs that try to explain social outcomes. More detail about 'getting the data' is given in other chapters of this book.[1] See also 6.1 **case-study research** and 7.3 **retroduction**.

Note

1. Some readers may wonder whether hypotheses are necessary for a research project. They are not. All projects need to have an initial research question, but this can be modified and is open to change. But having hypotheses or testing hypotheses is not well suited to all projects. The protocol in Box 3 is offered mainly to help those who wish to study through the use of hypotheses, so that they can do it well. 'Doing it well' may involve asking hard questions about where data are coming from, what causal processes or structures could have caused the data to appear as they do, whether the hypothesis is refuted, and so on. Data collection does not exhaust the research process, and in itself the data cannot guarantee good research.

1.5 Sampling

Sampling is a cost-saving way to create sets of cases. For instance, you might sample households randomly in a city, then select individuals non-randomly from those households. If you were doing a study of young people you might choose only those between the ages of 16 and 30, for example. The two main methods of sampling, non-random and random, can be merged only at the cost that the result is usually considered to be non-random. However, non-random sampling plays an important role in so much social research that I propose to discuss it first. Then when describing random sampling the possibility of multi-stage sampling makes a lot of sense. The chapter concludes by suggesting ways of using weights if the sampling scheme does not suit the population of cases that you wish the data to represent.

Non-random selection of cases cannot be called sampling at all. Instead we use the word 'selection' and a number of strategic decisions are normally made. For instance, one may decide *a priori* how many cases are desired, what the geographic area is, what unit is to be considered a case, and whether cases are to be seen in their nested location within larger units.

The first method of non-random case selection is snowball sampling. Snowball sampling refers to extending the network of known cases outward using contacts offered by those within the existing sample. Snowball sampling for small studies may stay within a tight geographic area. Suppose you were studying skiing accidents. You might pick up some cases from an indoor skiing facility, then ask them to name some other skiers. Using the phone or email, you would try to contact these skiers and ask them, among other things, for more contacts. During this process you would normally make a list containing basic details about each case (person, skier). If you decided to use the 'accident' as the case instead of the 'person', then you would amend the notes to suit the characteristics of the new case unit. Accidents are usually nested within skiers. In a few cases a skier may give you contact details of another skier involved in the same accident. By creating an *N*-to-*N* relationship, with *N* representing any number above 1, this makes a non-nested relation between skiers and accidents. Your task in qualitative research is usually to focus on one or the other of these. For survey research it might be easier to insist on a 1-to-*N* relationship, hence a nested relationship, in which each skier can have multiple accidents but each accident is associated with just one of the skier cases. Snowball sampling does not give a random representation of the population. Indeed, the whole population's characteristics (and location) need not even be known.

An example of snowball sampling (Eldar-Avidan et al., 2009) shows that in selecting 22 respondents it is possible to claim to have touched upon a wide variety of groups in society. The small-scale study by Eldar-Avidan et al., was focused on adults whose parents had been divorced while they were young. On the design of the sample they write:

> Twenty-two Israeli young adults participated in this study, aged 20–25 years (average: 23), whose parents divorced before they were 18. … Participants were chosen to ensure variability and sampling was terminated when saturation was reached …. Heterogeneity was achieved in regard to age (within the predefined age group), age at the time of divorce, time since the divorce, education, occupation, parents' current marital status, number of siblings, and place of residence. Thus some came from big cities, while others were from smaller towns, villages, or the kibbutz (communal settlements); some were still doing their mandatory military service, while others were working, studying, or both. Some participants were referred ('snowball') by colleagues or other young people. (Eldar-Avidan et al., 2009: 2)

The claim being made here is not that the sample is representative. For just $N = 22$ there can only be one or two from several of the groups that are named. It would be hard to justify claiming that this is a representative sample in the usual statistical sense of enabling inferences from sample to population. Instead, however, behind this sampling strategy is a mild assumption that most young people in Israel of the age group 20–25 whose parents were divorced would exhibit some homogeneity on some of the relevant characteristics. However, the authors are open to diversity; in the findings, there is more than a hint that tapping into a subgroup, such as rural Israeli residents, with just a handful of cases will give interesting and perhaps even representative information about them. The paper rightly avoids making generalisations across broad swathes of Israeli people. Instead most of the general statements in the paper are simply about the 22 young people interviewed. Like most grounded theory, the paper is highly concrete and the findings are specific to the selected cases chosen.

In the conclusions of the study, to some extent a distinction is made among three types of young people: those left more vulnerable after divorce, those more resilient after divorce, and those who are survivors of divorce. The results of this study are carefully couched in grounded theory as a method of data gathering and analysis. In grounded theory, purposive stepwise sampling goes hand-in-hand with developing a theory about outcomes. In this instance the stepwise sampling included some snowballing and some purposeful deviation into contrasting groups. The word 'saturated' is used in grounded theory when the sample selection is considered to be complete and the theory development has firmly started, but data are still being collected and analysed.

Snowball sampling can also be done via the internet. Internet groups and online sites such as dating sites offer additional opportunities for qualitative sample selection. Another form of qualitative non-random sample selection, quota sampling, depends upon making some decisions about the types of respondents that are wanted, making a grid of basic characteristics and distributing the desired sample size among them, and then going out to a specific area (or the internet) and finding people or other cases of the desired types. Quota sampling is often confused with random sampling by the less initiated. It is critical to realise that when filling a quota – from the street, from among students in a classroom, from among visitors to discussion groups or wherever – there is a strong tendency for a bias in sample selection to arise from among the cases available. This bias will differ for different researchers; to give two examples, young women might avoid tall older men in one scenario, while researchers of one linguistic group such as French might avoid speakers of another language group in some other scenario. The quota table might specify what sex and age group to choose, and what occupational category to accept into each quota. If it specifies the areas that are at risk of bias, then the tendency to bias can be compensated for. If it does not, then this bias risk remains.

Random sampling, on the other hand, attempts to avoid bias by asking the enumerators or researchers to select specific cases from a given list of available cases. An obvious example is to use the electoral register (which contains adults who are registered to vote in a specific geographic area, and their addresses), selecting a certain random sample of pages and then a random sample from each page. One can generate random numbers in Microsoft Excel with =RAND(). For instance, using Excel, one can generate 60 random numbers from 1 to 12 using the formula =INT(RAND()*12+1. (The function starts at 0 by default so needs 1 to be added to get positive integers.) If the register has 96 names per page, one may choose names by counting downward using the first random number to a name, then to the next, e.g. down 6 lines and then down 4 lines. For each page, on average, 8 names will be chosen. If 7 pages had been random selected to begin with, then 7 × 8 = 56 would be the achieved sample size (nearly 60). By adjusting the formula, a sample of any target size can be achieved. From the 56 individuals, only a very small handful will be living together in the same household. This sampling frame is thus said to be a list of individuals, not a list of households. If the list is randomly ordered, then a simpler method called systematic sampling can produce a random sample. Take every *n*th name, e.g. every 12th name, on each selected page.

After generating the sample frame, one calculates the sampling fraction as n/N, where n is the number in the whole sample and N is the number in the list which is being taken as the population. The sampling fraction is irrelevant to the degree of statistical significance you will have in calculating statistics from the study. What really matters is sampling non-response (where cases are not available, or individuals refuse to answer) and the absolute size of the final sample. The non-response rate is the number of refusals divided by the total *n* in the desired sample. We need to denote the actual sample size by *n′*, indicating that *n′* is less than *n* by the number of refusals, and the effective sampling fraction is *n′*/*N* rather than *n*/*N*. In social science a sampling fraction such as 85% is considered good, and 95% excellent, while in market research and business studies some lower sampling fractions are considered acceptable. Postal surveys and those projects where respondents are considered to come from a homogeneous population are areas where a lower sampling fraction is considered more acceptable. If bias is suspected in the non-response rate, the rate should be high. If bias is not considered a problem, perhaps because of homogeneity, then the non-response rate can be allowed to be low. Homogeneity of the cases might exist, for example, if all taste-buds worked the same general way and consumers were being tested for how sweet they thought particular foods were. The homogeneity that matters is only on all the characteristics that matter for a given study. If, instead, one is looking at consumer preferences for high- and low-alcohol drinks, and the level of alcohol is not easily perceived but instead is being guessed at using labelling clues and past experience, then we would have to assume a heterogeneous population and try to get a sample with a high response rate and no researcher selection biases.

Cluster sampling refers to choosing cases from particular places or groups which are named and listed prior to setting up the next stage of sampling. If there were 30 bowling clubs in a region, and the cases desired were Saturday night bowlers, then one could choose five bowling clubs randomly and move to a second-stage sampling strategy on Saturday nights of taking all the bowlers on three lanes at random in each club once per hour from 7 to 11 p.m. In the cluster sampling strategy, the next stage can instead cover the entire population within each cluster, e.g. all bowlers present on Saturday night over a three-week period. Cluster sampling runs the risk of missing heterogeneity in the clusters not chosen. It also depends upon some prior information such as a list of bowling clubs. Cluster sampling is usually considered to cause wider confidence intervals when sophisticated statistical software is allowed to integrate the information about clustering with the rest of the detailed survey data. However, for some studies cluster sampling is a highly efficient way to target the visits of researchers to specific areas, and for this cost-saving reason it is desirable even if it leads to a risk of poor representation of a geographic region overall.

Stratified sampling is more complicated. To stratify a sample, one decides first on the overall sample size, then breaks this *n* down into subgroups. Prior information about the prevalence of certain characteristics, relevant to the study, in each subgroup is used to decide on the percentage of cases which are to come from each stratum. The strata can be of different sizes, and a simple mode of stratified sampling is to use the inverse of the proportion in the group to set the proportion of *n* that will be in that stratum. Small strata thus get better representation, and large strata worse representation, than in simple random sampling. It can be proved that if the strata are set up well to encourage large strata only where there is case-wise homogeneity on relevant characteristics, then stratified sampling will tend to give a better representation than simple random sampling over the chosen population. However, there are often difficulties getting the relevant information. Once the strata are set up, it is also possible to manipulate the sample sizes and make a post-sampling weighting adjustment (see 5.8 **survey weights**) to give unbiased averages and other statistical results for the whole *n* overall. The use of post-stratification survey weights can also be a way to adjust for non-response and leads to design-based statistical inference. Cluster sampling can be used in conjunction with stratified sampling. Just when cluster sampling has reduced the overall accuracy of a survey sample, stratification within each cluster can attempt to increase it and thus improve accuracy again. The mixing of cluster sampling with stratified sampling is usually called multi-stage sampling.

If multi-stage sampling includes a non-random stage, such as snowballing, then the sampling method overall is not considered random. If it is not random, then generalisations based on the sample *n* should not be considered to apply to the wider population from which the sample was drawn. Care needs to be taken not to confuse these basic rules if random sampling is considered desirable. In general, random sampling is more expensive than non-random case selection. However, the overall ambitiousness of a study – its geographic and social extent – is an easy factor to vary if cost savings are needed. In the case of Israel, for example, one could have had an urban-only study with a higher number of cases for the same cost.

If non-random sampling is used, including carefully manipulated sampling strategies such as cluster sampling and stratified sampling, then a weighted average of the data in the rows can be used to generate an estimate of the average overall among a sample as a representation of the average in the population. The error in this estimate will tend to be increased where cluster sampling is used, but on the other hand error is decreased where stratified sampling is used. The researcher who has either cluster or stratified sampling may choose to use post-stratification weights. These weights imply design-based inference. Furthermore, those who have any noticeable non-response rate may create and apply post-sampling weights by comparing the actual sample with a gold-standard data source, such as a census or a full hospital register. Post-sampling weights are usually called post-stratification weighting, but there are three sub-kinds of these weights. One is an identifier for each cluster; the second is an identifier for each stratum and furthermore a weight attached to the stratum that compensates for higher or lower sampling in that one relative to overall; and thirdly the non-response weighting adjustments. All the weights average out to 1 and they can be multiplied by each other. The result is the overall weight given to each case for design-based inference. For example, this would produce the mean of age for the sample with a 95% confidence interval.

Model-based inference is slightly different from design-based inference. A researcher who has access to enough variables that relate either directly or indirectly to the main axes of sampling bias, and who wants to use regression or another multivariate modelling technique, can avoid using post-stratification weights. For example, the data could have a large bulk of students, and then a variable 'student' is put into the model and variation according to this is thus controlled for. The remaining variation in an outcome is examined for how it varies with some other variable – perhaps body size, or previous training. Cleaning out underlying sources of variation in a model-based framework is most frequently done in econometric and panel-data contexts (Greene, 2003; Wooldridge, 2002). It may be comforting to know that it is possible to examine experimental data or other survey data without random sampling. In general, it is best to do so only after making explicit several assumptions about the nature and diversity of the cases and the population. There must also be sufficient cases in the sample *n*. Another approach would be to make statements that are purely descriptive of the actual sample, but this interpretive stance would not involve inference. Inference is defined as making a valid statement about a population based on a random sample from that population.

Further Reading for Part 1

Data gathering and data analysis are difficult to disentangle. The works I would suggest for further reading also link up these tasks, so that the researcher cannot just automate a data-collection process or follow a rigid rule-book. Instead, people tend to think about how they might attack a problem, develop a research question around that, plan the data collection and data analysis, embark on data collection – perhaps with a pilot study – and then revisit the earlier stages before completing the whole data-collection stage.

Blaikie (2000) and Danermark et al. (2002) strongly support this overall approach which integrates data collection with a holistic well-managed research project. Blaikie (2000) walks you through all the stages of research design. Danermark et al., (2002) have provided a protocol for research which gives an overarching map of the contours of progress through a project. A useful book of comparable scope, with more technical detail and less on the methodology side, is DeVaus (2001). Layder (1993) takes up the issues in a very straightforward language, suitable for undergraduates and the newest entrants to the research scene. This lucid book helps the reader choose methods suited to their particular kinds of research topics.

For those working in Third World or comparative contexts, Laws (2003) provides a book on data collection and methods. In this general area there are two full-length sources of more concrete guidance, both very lucid and helpful. One is written entirely by one author (Mikkelsen, 1995, 2005). She takes into account the difficulty of separating development research from development practice. Although she covers quantitative and qualitative primary data collection practice in some detail, she also has a wonderful section on participatory data collection and encouragement to mix methods. An edited volume on development research as a form of practice is Thomas et al. (1998).

Participation in research is explained in technical detail by Mikkelsen (1995). I recommend the edited volume by Carroll et al. (2004) for anyone wanting to do engaged research, or participatory or action research in any social context. The overview by Flyvbjerg (2001) says much less about practical methods, but it does help by making sense of the whole scene, with researchers being purposive people participating in a social and political landscape.

See also Barnett (2002) on sampling.