Research Methods
Sample Lecture Pack
Code: REM020

Postgraduate Study in
Educational and Social Research by Distance Learning

This is an extract from a lecture pack for a module offered as part of the University of London International Programmes Master of Research (MRes) in Educational and Social Research. Materials for this degree programme are developed by academics at the Institute of Education of the University of London.
For more information, see: www.londoninternational.ac.uk
Module 2 Research Methods

Module introduction

by Will Gibson

Theory, method and practicalities

There can be no sensible method to study a domain, unless one also has a theory of what the domain is. (Gee, 1999: 5)

Methods of data collection are of course a central concern for social researchers; figuring out how to use and conduct interviews, questionnaires, observations, and documentary research is fundamental to the enterprise of becoming a social researcher. Like all areas of social research the discussion of method is not easy to hold in isolation - as the above quote from Gee highlights, decisions about method are always tied up with decisions about theory, be that in terms of specific epistemological positions like interpretivism or positivism, or more localized theoretical orientations to the world, like the use of a specific perspective (such as sociology), theory (for example ascetism and work ethics) or theorists (like Max Weber). When thinking about methods, there are very many practical matters to attend to - like 'How many interviews are enough?', 'What is the best way to write this question in my questionnaire?', 'How do I choose who to include in my study?', or 'Which documents should I collect and which ones should I ignore?' - but these very often bring you back to consider more abstract issues, like 'What is my view of the interpretive character of social life?' or 'What is my stance on generalization?' Theory, in one form or another, then, can be an important resource in helping you make these choices.

For example, say a researcher is interested in community outreach projects run by a particular charity that provides sex education in Africa. Perhaps that researcher is especially interested in the ways that the social organization of the communities in which such programmes are run may provide alternatives to the formal classroom as a means of disseminating the information that such projects aim to distribute. Perhaps the researcher has a particular interest in the concept of 'power' that has developed through a reading of particular social theorists (through reading scholars in the field of history and postcolonialism, for example). The researcher's interest may be in particular forms of social organization - perhaps patriarchal and/or tribal societies - which may lead him or her to select specific sites for their study. Similarly, the nature of the settings may rule out certain methods of enquiry: perhaps questionnaires are not feasible because of low levels of literacy (literacy in the form that is needed, anyway). Document analysis may be hard too if the types of documents that the researcher has the skills to interpret (like reports, diaries, notes and so on) are not typically used in the cultures being researched. Observations might well be problematic: it could be difficult to be accepted as an insider to such 'isolated' communities, which could leave interviews as the obvious (and only) 'choice' of data collection method. What this example is intended to highlight is that the setting itself and the nature of a researcher's interest in it may well place logical limits on the choices over which methods to use.
Ethics

The ethical implications of specific research designs will also be a central aspect of the ways in which the final decisions about methods are made. Medical and educational research are notoriously riddled with ethical problems and restrictions. The role of researchers is to work through the specific ethical issues presented within a given project in relation to the practical interests of their research programme. There remains considerable debate within the research field about what constitutes good ethical practice, with different organizations orienting to quite distinct ethical procedures. The challenge for researchers is to design and conduct research that both fits the specifications outlined in rather abstract ethical guidelines and meets the specific challenges brought by research in a particular empirical field, in a way that produces useful data.

Returning to the example we gave above, let's assume that our researcher is male. Maybe there is nothing in a particular set of ethical guidelines that says that the researcher can't interview the wives of community elders, but perhaps there are specific customs about the interactions of men and women that would have implications for how the women were treated by the community subsequently, were the researcher to do so. Perhaps, then, the researcher decides not to interview them, but to just hang around in communal gatherings and to try and steer conversations around to relevant topics. The researcher might need to be quite creative to find a way to get the data that he wants in a way that can be defended somehow as ethical (although it is of course very debatable whether the strategy outlined is in fact particularly ethical).

Research questions and research methods

Another central referent in decisions about method is the specific research question/problem/issue being addressed. An extremely useful way of developing research ideas, particularly in the early stages of research, is to think through the ways in which different research methods may be used to address the particular issue being studied. In the abstract, any research method can be used to research any given problem, but each method is likely to lead to, or imply, a particular kind of focus. For example, one of my interests is in the ways in which doctors explain medical conditions to patients in the context of one-to-one consultations. I might research this by examining documents such as diary accounts from patients or doctors of their consultations, or by analysing educational materials that teach doctors what to do in these types of settings. Through methods of this kind I might be able to do an analysis of, say, the changing nature of consultation practices through time, maybe comparing practices in the late 1800s with contemporary approaches. The kinds of documents that I could find would very much define my specific question - if I could find only detailed notes from a particular doctor in the north-east of England dating back to, say 1887, I might decide to simply do a case study of that surgery. If I found quite a variety of evidence in that time period, however, I might use a very different design, perhaps comparing practices in different parts of the country.

I could equally use questionnaires to do my research. This would of course mean examining contemporary practice rather than historical ones, and it might lead me to ask how differences in gender, age, level of experience, training methods or any other variables might influence consultation practices in terms of, say, the length of time that consultations last, the types of technologies that are used in consultations, or the depth of information that is provided by doctors. This sort of
approach would, in principle, allow me to reach quite large numbers of people and perhaps to compare practice in a number of different socio-cultural-economic areas (maybe even in different countries).

Alternatively, I could use interviews of one type or another to talk to doctors or patients in order to find out the details of their attitudes or experience. I might ask them to give me examples of particular experiences they had had, or I could pose hypothetical scenarios and ask them to comment on them. In this way I could probably get more detail than I could with questionnaires, but I would be much more restricted in terms of the numbers of people I could talk to. I might decide to observe some consultation sessions so that I could see in even more detail what goes on. I would then be able to look at the specific language that is used in consultation settings, or the ways in which doctors and patients negotiate what gets talked about within a particular session. I might decide, though, that there has actually been quite a lot of research done in this area and that what would be really useful is to build up a detailed picture of this research by doing a systematic review. I might want to compare the types of findings that different studies have generated, and perhaps to produce an evaluation of consultation protocols on this basis.

The choices that a researcher makes about research methods will define the parameters of the research by creating some restrictions and opening up some possibilities for the specific ways in which questions can be asked and the types of claims that can be subsequently made.

Preferences for particular methods

Researchers very often do have preferences for particular methods. There are good reasons for this, of course; researchers very often orient quite strongly to particular sets of interest, and these interests guide them towards certain types of data. In a classic sociological article, Benney and Hughes (1956) suggest that different subject areas - like history, sociology, psychology and so on - are characterized by preferences for particular methods.

> The several branches of social study are distinguished from one another perhaps more by their predilection for certain kinds of data and certain instruments for digging them up than by their logic.
> (Benney and Hughes, 1956: 137)

While in general terms this may have some element of truth to it, the matter is really much more complex than this, because each subject is made up of sub-disciplines, each of which is likely to have its own preferences. So, for example, symbolic interactionists (as a branch of sociology) will rarely be seen handing out questionnaires, because their interests in the nature of meaning construction are not easily reconciled with the closed categories of a questionnaire. Other sociologists, however - particularly those with more quantifiable interests - will rely very much on this method of data gathering because it will enable them to collect information from a very broad sample of the population and to answer more generalized questions.

Preferences for particular methods, however, do not just come from theoretical or empirical specializations, but also from the pragmatics of research. For example, research in social policy within the United Kingdom is strongly quantitative because figures are politicians' preferred mode of representing research findings. Anyone wishing to get funding for government research therefore has more
chance of success if they adopt techniques that result in enumerated data than, say, broad descriptions or narrative interview. Very often, funding bodies have strict limits on the amount of money that can be applied for as a research grant and, as with every area of life, money is a serious consideration in making any plans.

While it is certainly not always the case that researchers operate with such preferences or pragmatic interests, it is nonetheless a little idealistic to think of research as always involving a free choice between methods. The interests of the researcher, the parameters of time, money, personnel and other resources, and the preferences of the funding body may well push researchers in particular directions. That said, there is a strong analytic benefit in thinking through the ways that a given project may be explored by different methods; it helps you to think about how different sorts of data (which is what methods are all about in the end) may relate to, or help you to frame, a given research problem.

In your work for this module you will think about the specific contingencies that researchers face in using all of the methods mentioned above. In this way, you will gain a broad understanding of the central issues involved in using a variety of methods, and get experience of actually using these methods in the context of real-world research. As ever, there is a strong benefit to be gained from using and developing your own research interests in the context of this course.

**Literature about research methods**

As you will no doubt have discovered for yourself, there is a huge amount of literature dealing with research methods. This is both a help and a hindrance, because while it is great to have such a wealth of materials at your disposal it can be hard to know where to start. The literature selected as key and supplementary reading in this module has been chosen specifically because it deals with central issues, because it comes from prominent authors, or because it is a particularly influential piece of writing. There is no one key text for this module, but each unit has outlined key readings to help you through the topics. The textbook by Brown and Dowling (1998) provided for this course is, however, a very good resource for many of the issues that will be discussed in this module.

You are encouraged to read as widely as possible, not just in research methods textbooks but to also look at empirical studies that involve real-world applications of the methods you will be exploring. You should develop your own personalized library of literature, and keep good records of the references you collect, including detailed summaries of the arguments, selected quotations, keywords, and outlines of methods, findings and any other particularly relevant features. Being able to refer to and critique the literature (both empirical studies and more abstract discussions of research methods) is a key aspect of the assessment criteria for this module.
Module content and structure

The module is separated into four sections, with each section corresponding to a particular online activity. The units are grouped as follows:

| Section 1 Asking questions | Unit 1 Questionnaires  
|                           | Unit 2 Interviews  
| Section 2 Observing people | Unit 3 Observation methods  
| Section 3 Researching online | Unit 4 Internet research  
| Section 4 Research and documents | Unit 5 Systematic reviews  
|                             | Unit 6 Documentary research  

Your instructions for using the unit material and the assigned readings are provided online along with full details of the activities, including their start and finish dates. To reiterate, this is a VLE-centred course, so please regard these units and their accompanying readings as resources for undertaking the activities rather than as forming the main content of the course.

**Learning outcomes**

By the end of the module you should be able to:

• critically evaluate a questionnaire in a manner that shows an advanced understanding of the issues of questionnaire design

• critically examine the debates surrounding interview research, and relate these debates to the construction of a real-world research project

• evaluate observation-based research, using hypothetical and real-world scenarios, showing advanced knowledge of relevant methodological debates

• outline and effectively orient towards key debates in internet-based research

• formulate a systematic review in an original area which explores a well-specified research question

• critically analyse the ways in which documents may be used to frame or explore a research question

• construct an original research strategy that appropriately combines a range of research methods and shows an advanced knowledge of the debates surrounding their use.
References


Module 2 Research Methods

Section 1 Asking questions

Unit 1 Questionnaires

by Jane Hurry

This unit will explore some of the key issues involved in questionnaire research, including the ways in which conceptual issues, such as reliability and validity, can be taken forward as practical strategies of research design. You will also look at a number of approaches to creating samples and other general design issues in constructing questionnaires. Through this unit, you should get a very clear idea of the central debates within questionnaire methodology, and develop the tools necessary to undertake questionnaire research. We begin this discussion with some exploration of the theoretical issues underpinning questionnaires.

After studying this unit you should be able to:

• produce a questionnaire with an awareness of some of the key debates in questionnaire construction
• critically evaluate the concepts of reliability and validity
• describe and compare a variety of approaches to sampling.

Key readings for this unit

• ‘Surveys and Sampling’ in Research Methods in Psychology by Patrick Sturgis (2003)

A theoretical framework

Questionnaires are typically seen as reflecting a 'positivistic' research orientation, seeking to capture an 'objective', 'real' world 'out there'. This approach is generally defined by a number of assumptions and these are listed below. This list (based on Foddy, 1993: 13) is useful not merely as an academic exercise but as a sort of check list for those using this approach. If you are using a questionnaire format or even, to some extent, a semi-structured interview format you should run through the list to see if you have considered these issues:

• The researcher has clearly defined the topic about which information is required.
• Respondents have the information that the researcher requires.

• Respondents are able to access the required information under the conditions of the research situation.

• Respondents can understand each question as the researcher intends it to be understood.

• Respondents are willing (or at least can be motivated) to give the required information to the researcher.

• The answers that respondents give to a particular question are more valid if the researcher has not suggested them to the respondents.

• The research situation *per se* does not influence the nature of the answers given by respondents.

• The process of answering questions *per se* does not change the respondents’ beliefs, opinions, habits etc.

• The answers that different respondents give to a particular question can be meaningfully compared with one another.

The aim of questionnaires is to achieve common meaning through the exchange of questions and answers; this is achieved by asking questions in the simplest form possible. These questions are designed to remain faithful to the conceptual intentions of the researcher while expressing the issues in a clear way. To these ends, the major effort of this method of research is in ensuring that the questionnaire is as clear, precise, unambiguous and intelligible as possible and not, for example, leading, hypothetical, embarrassing or memory defeating.

In contrast, qualitative field researchers (such as ethnographers) adopt a phenomenological or 'subjectivist' point of view, which highlights the process of interpretation that is present in formulating and responding to questions. In this approach, different 'respondents' will often have quite different interpretations of the same question, and the answers given by a particular respondent to a question in one social situation may be quite different from the answers given by the same respondent in another situation.

The most common criticism of questionnaires is that their preset response categories determine the way the respondent can answer the question. Put another way, it is argued that providing sets of response options may cause respondents to give answers that they would not think of if they had to create answers themselves. In contrast, the major criticism made of qualitative researchers is that their observations are open to subjective interpretation and reporting: that observations can be selectively reported to give a distorted picture, that interpretations typically have very low levels of reliability, and that it can therefore be difficult to replicate findings. Also, even though qualitative field researchers honestly strive to 'absorb' the culture of the members of the social group under investigation, in the final analysis they have either to infer or to guess the nature of the cultural elements. Researchers need to find their own way in this diffuse field of research practices. If you are using questionnaires, it is important to be clear about the epistemological assumptions that you are making and be aware that they differ from assumptions underlying other research paradigms. We will now turn to some of the more practical issues involved in questionnaire design and construction.
The key issue: comparability of answers

Questionnaires can be thought of as questions that pose limits on the range of answers that can be given. If a question-answer sequence is to make sense, the question must be understood by the respondent in the way intended by the researcher. In addition, the answer given by the respondent must be understood by the researcher in the way intended by the respondent. The four essential steps in a successful question-answer sequence are as follows:

1. The researcher must be clear about the nature of the information required and encode a request for this information.
2. The respondent must decode this request in the way the researcher intends it to be decoded.
3. The respondent must encode an answer that contains the information the researcher has requested.
4. The researcher must decode the answer as the respondent intended it to be decoded.

It is part of the process of designing questionnaires to attempt to ensure that these steps can be successfully followed.

Advantages and disadvantages of questionnaires

Questionnaires offer a number of advantages:

• They can reach a large number of people quite easily and economically (especially postal questionnaires). That might seem a trivial advantage, but the time savings against travelling to see participants can be substantial.

• They are relatively easy to analyse.

• They provide quantifiable answers.

Questionnaires also have disadvantages that relate closely to the advantages:

• Although postal questionnaires are easy and efficient, it is much harder to check whether the answers are genuine.

• Postal questionnaires can have a fairly low response rate (less than 20 per cent).

• It is not unusual to find incomplete questionnaires as a result of deliberate and accidental omissions.

• The structure of the questionnaire can too easily reflect the way the researcher sees the issues, both in terms of questions chosen and the way the questions (and possible answers) are phrased.

Stages in designing a questionnaire

Designing questionnaires from scratch is hard work and for something to be considered methodologically rigorous it is necessary to address reliability and
validity. For this reason, where possible, it is helpful to use existing questionnaires. This has the added advantage of enabling comparison with the work of others. To find existing questionnaires, refer to relevant studies in your area of interest, search web sites, specialized books on test collections, etc. Where questionnaires have been extensively used, they are sometimes referred to as standardized tests, such as the IQ test (more specifically the WAIS - Wechsler Adult Intelligence Scale). Such tests may have population norms so that you can compare your respondents with the population as a whole. Sometimes no adequate measure exists and then there is nothing for it but to design your own. Whichever approach is adopted, the first stage is to clarify the research questions and the purpose of using a questionnaire. Why do you want to know the answer to a particular question? What is its relevance to your research questions?

Following this, the next stages are:

1 Identify the target population.
   The people who will complete the questionnaire have an impact on the way a questionnaire is designed. If you are using an existing questionnaire to collect data, it is important to ask whether or not it was developed using similar respondents. If not, it may be necessary to think about or demonstrate whether it is reliable and valid (these terms are discussed in greater detail below) for your respondents. If a questionnaire is being developed from scratch, it should be appropriate in terms of language, response categories, etc. for the respondents who will complete it. For example, you may need to use different language for children compared to adults.

2 Identify concepts that need measuring.
   For example, if a researcher is interested in looking at the effect of social class on children's progress in school, the concepts of interest might be 'social class' and 'school progress'. Now, this provides some basic information to enable you to search for existing measures, questionnaires and tests. If measures for these concepts do not exist, then it is necessary to:
   • develop indicators - for example, if social class is the concept, key indicators might be: occupation, income, education level, house ownership, number of cattle, inside toilet, running water, mother and father's level of education, and so on
   • develop question items
   • decide on the content of questions
   • decide on format and wording of questions
   • decide on the presentation and layout of the questionnaire
   • formulate a coding schedule (if appropriate)
   • pilot and refine the questionnaire.

Here are some useful tips to think about while working through these stages.

• It is best not to formulate specific questions until you have thought through the research question.

• Write the research questions down and keep them handy while formulating specific questions.

• Keep asking 'Why do I want to know this?"
The chapters from de Vaus (1996) that have been selected for you to read in conjunction with this unit further discuss the details of questionnaire construction and piloting.

**Rating scales**

De Vaus refers to the wide range of types of questions and rating scales and some examples are given below:

**Numerical rating scale**

<table>
<thead>
<tr>
<th>How useful would you think a family planning centre in your area would be? (please circle)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>Not at all useful</td>
</tr>
</tbody>
</table>

**Checklist**

Please tick the boxes beside the characteristics that best describe you as a person.

- Ambitious
- Athletic
- Cautious
- Good-looking
- Moody
- Happy
- Obliging
- Highly strung
- Poised

**Likert scales**

A local family planning centre would be useful for my community (please tick):

<table>
<thead>
<tr>
<th>DISAGREE</th>
<th>DISAGREE</th>
<th>NEUTRAL</th>
<th>AGREE</th>
<th>AGREE</th>
</tr>
</thead>
<tbody>
<tr>
<td>strongly</td>
<td>strongly</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Ranking scales**

Rank order the following issues in terms of how important they are to you when you decide to vote (1 would be the most important and 5 the least important).

- Policies to reduce unemployment
- Improving the environment
- Spending more money on education
- Getting tough on crime
- Reducing taxation
Visual analogue scale (also ‘semantic differential’ rating scale)

Please indicate how helpful your employer has been in enabling you to develop your career by putting a mark on the following scale.

Not at all helpful → Extremely helpful

Dichotomous questions

What is your sex (gender)?

Male □ Female □

Do you smoke cigarettes?

Yes □ No □

Multiple nominal responses

Please indicate your current marital status by ticking the appropriate box.

Married/cohabiting □
Never married □
Widowed □
Separated □
Divorced □
Other □

Multiple ordinal responses

Please indicate how often you attend church or a place of worship by ticking the appropriate box.

At least weekly □
Two or three times a month □
About once a month □
Once every three months □
Almost never □
Never □

Collecting background/demographic data

It is very useful to collect background data on the respondents to questionnaires so that you can learn about possible relations between their social characteristics and their responses. It is important to think carefully about the background data that are likely to be useful in your study: sex, age, social class (whose social class and how can it be assessed?), education, ethnicity, household composition, marital status, size of institution, characteristics of institution etc. Not all data are
useful, and only data that are useful should be collected. If a questionnaire becomes too long, there is a chance that the level of response will be reduced.

**Sampling**

One way of finding out about a group of people is to collect information from everyone in the group. For large groups of people this is prohibitively expensive and impractical. The alternative is to collect information from only some people in the group in such a way that their responses and characteristics reflect those of the group from which they are drawn. This procedure is much cheaper, faster and easier than surveying all members of a group. This is the principle of sampling.

If you look up a chapter headed 'Sampling' in a research methods textbook, you will almost invariably read about sampling for survey designs, usually large-scale surveys. However, all designs require sampling strategies, experiment, case study, ethnography, historical investigation, etc. The most technical sampling requirements are for survey designs and this is what we will consider first.

**Surveys**

A few technical terms:

- A **census** is obtained by collecting information about each member of a group.
- All members of a group are called a **population**.
- A **sample** is obtained by collecting information about only some members of the population.
- Samples can reflect the population from which they are drawn with varying degrees of accuracy. A sample which accurately reflects its population is called a **representative** sample.

**Major principles underlying survey sampling**

1. Avoid bias.
2. Achieve maximum precision for a given outlay of resources.

You want to be able to generalize from your sample to the population sampled. If there is bias in the sample, by definition this will be problematic. On the other hand, the decision to sample rather than to collect information from every member of the population recognizes the need to take available resources into account.

It is essential that you have good understanding of the population from which you are sampling and that you make this explicit. There is nothing wrong with sampling from a population of, for example, only women, or only people of a certain ethnicity, but you must understand the limits this places on generalization.

**Random or probability sampling**

Random sampling gives every member of a population an equal opportunity to be sampled.
The importance of randomness in the selection procedure cannot be over-emphasised. It is an essential part of the protection against selection bias. (Moser and Kalton, 1971: 82)

The whole theoretical framework of much statistical analysis rests on the assumption of random selection.

**Types of random sample**

**Simple random sampling**
- The lottery method (numbers drawn out of a hat, for example).
- Using random number tables (see the references in Moser and Kalton, 1971).
- Using computers - most quantitative software packages allow you to select a random sample directly.

**Systematic sampling**

This involves picking subjects at a fixed frequency from an ordered list of the population (for example, in an alphabetical list choosing every third child on a school register). It may be more difficult to identify such lists for certain populations in countries where national resources available for some types of administration are small.

The definition of randomness relates to the mode of selection, not to the resultant sample. It is possible to select a sample randomly but to find that you have drawn an unrepresentative sample. For example, you could randomly select a group of children from a mixed school and find that they were all boys, or that a much higher proportion of your sample were boys than the proportion in the school as a whole. This is one of the potential drawbacks of this kind of random sample.

One can avoid this problem to some extent by using a stratified sample (still random).

**Stratified samples**

In a stratified sample, before any selection takes place, the population is divided into strata. Then, a random sample is selected within each stratum.

> If the sampling fraction is the same for every stratum, this procedure is almost certain to be an improvement on a simple random sample because it makes sure that the different strata in the population (sexes, age groups, regions, town types and the like) are correctly represented in the sample. (Moser and Kalton, 1971: 85)

In the example given above of children attending a mixed school, suppose that 200 of the children are boys and 100 girls. In a **proportionate stratified sample** where you want to select 1 in 10 children, randomly selecting from the boys and the girls separately would ensure 20 boys and 10 girls in the sample.

Moser and Kalton (1971) give a very interesting discussion of the mathematical proof of the increased precision that the stratified sample has over the simple
random sample. Of course, using a stratified sample implies that you have
detailed information for the population you are sampling from, and that you know
what strata are likely to be important from the point of view of bias in your own
study. There is no gain in precision from stratifying by a factor unrelated to the
subject of the survey. You could stratify children by whether they are short or tall,
whether or not they like vegetables, or whether they watch cartoons on TV, but if
your research questions do not relate to any of these factors, you will not have
improved the representativeness of your sample for your purposes.

Sometimes you may want to sample proportionately more from one strata than
another - disproportionate stratified sampling. The most likely reason for this
is that you want to ensure that minority groups that you are particularly interested
in are represented in sufficiently large numbers in your sample. For example, if
you are doing a study of the difference between the reading progress of English
white children and Afro-Caribbean children, and you are drawing your sample
from London schools, you will need to ensure that you have sufficient numbers of
Afro-Caribbean children to make comparison possible.

The question that research students often ask, 'What sample size do I need for
my study?', is relevant here. The discussion in the section on sample size will
give some clues as to how to address it but, as a general rule, it is good to
remember that if you have a very small number in a particular group you will not
be able to say much about it, even if the overall sample is huge.

Moser and Kalton (1971) highlight another use for disproportionate stratified
sampling. Where there is a greater variation within some strata it would be
sensible to take a larger sampling fraction in the more variable strata. Say you
are doing research about literacy practices in different schools. One group of
schools may be governed by a closely managed curriculum and teaching
strategy. You might expect practice in these schools to be homogeneous and so
you could build up an understanding of literacy practices with a fairly small
sample of questionnaires, observations or interviews. Another group of schools
may be much less controlled and literacy practices much more variable. Here it
would be helpful to have a larger sample.

Cluster sampling

A cluster sample is one where the sample is randomly drawn from clusters of the
population, for example children from six Islington schools (Islington is a borough
of London) where the population is Islington schoolchildren, or three different
London boroughs where the population is London. Some random sampling is
more random than others! Cluster sampling can introduce bias, even though
selection within the cluster sampling is random. The good thing about cluster
sampling is that it is frequently very convenient, and sometimes necessary in
practical terms. What you have to watch out for is intra-class correlation. In
other words, the members of a cluster are more like each other than they are like
random members of the larger population. For example, the children attending
one school are likely to be from similar social backgrounds and less like children
attending the school down the road.

Multi-level sampling

As a result of intra-class correlation it may be interesting to look at certain
variables both at the level of the individual (pupil, patient etc.) and at the level of
the cluster (the school, hospital etc.). If a researcher wants to look at school
effects or hospital effects you need to think of the cluster unit (like school or hospital) as the unit of sampling, and not the individuals that make up those clusters. This involves sampling a number of those cluster units (that is, a number of hospitals, schools etc.).

Multi-phase sampling

Multi-phase sampling is when you sample a population, collect a range of information, and then select again from this sample an even smaller number of subjects and collect further information. Usually this is done to allow for the collection of detailed information in the second, smaller sample. This is a popular and valuable technique. Information collected on the full sample can be used to improve the precision of the sub-sample, for example by developing strata.

Non-random sampling

Quota sampling

In quota sampling there is a target number set for particular types of people who need to be included to represent the population: for example, 20 housewives, 30 working women, 60 men aged between 20 and 40 years. The advantage of quota sampling is that it is cheaper than other sampling methods, and it can be undertaken without knowing much about the population being studied. However, the major disadvantage with quota sampling is that it is very prone to sample bias. This technique is not recommended, except in exceptional cases to someone doing an academic piece of work - see Moser and Kalton (1971) for more details.

Cross-sectional sampling

This involves samples drawn at one particular point in time.

Time series sampling

Samples drawn at more than one time point, with different groups sampled on each occasion.

Longitudinal/cohort studies

The same sample of people, schools etc. is followed up at more than one time point.

Experimental designs

Experimental research designs typically require the comparison of two or more samples. By comparing the effects of the 'experiment' or 'treatment' or 'intervention' of one group with another group that did not receive the same intervention (usually referred to as a control group), the researcher aims to identify the effect of the intervention. The comparison groups therefore need to be as similar as possible before the intervention, otherwise it is difficult to interpret any subsequent group differences unambiguously (this is often referred to as internal validity). This can be achieved through random assignment of one
sample to two or more comparison groups, but it is not always possible. Another method is to draw one sample and then match the second sample to the first.

The critical factor in sampling for experimental designs is to ensure that the two groups are similar. Because experiments typically require a high level of cooperation from the participants, it can be difficult to carry out surveys using rigorous representative sampling (it can be difficult to recruit participants). However, if the sample is not representative, the confidence with which one can generalize to a population is limited. In some research there is an underlying assumption that the theory being tested is universal and so variations in the overall sample make little difference. For example, you might assume that a technique of setting broken bones is applicable to all humans because people's bones are essentially the same. However, in many contexts this assumption would not apply, particularly in social research. In such circumstances, the overall sample may need to be drawn using principles of representativeness, more akin to case study design (more on this later in the unit) and sample characteristics should be presented in reporting and taken into consideration in interpreting the findings.

**Non-response bias**

Whatever the method of sampling being used, there will always be some participants who drop out of a study, or who refuse to participate. It is necessary to record every non-response and, where known, the reason for it. It is important for researchers to report their response rate and the breakdown of reasons for non-responses. There is always a residual concern that those people who do not respond differ in some systematic way from those who do, and that this systematic difference is in some way linked to the focus of the research.

There are ways of improving response rates. These include:

- making sure the participants are interested in the research
- being careful to plan the research so that participation does not inconvenience the participants (it might take place at a particularly awkward time of day)
- convincing the participants that you are a legitimate (and institutionally affiliated) researcher
- following up any non-responses with a letter, phone call, e-mail or personal visit.

There is cause for concern when a response rate drops below 75 per cent and the higher the response rate, the better. Sometimes a low response rate is genuinely unavoidable. In these situations, researchers need to try to understand how the sample compares to the population from which it is sampled in order to make suitable inferences about generalizability. Indeed, it is always a good idea to show how a sample compares to the population in terms of the key variables of interest.

**Negotiating access**

For some types of research, it is possible to approach possible respondents directly and ask them whether or not they will participate. On other occasions it is necessary to gain permission from a third party, either to collect data from
individuals or to gain access to data on individuals held by an institution. An example of the latter case in educational research is where a researcher wants to speak to or test children. In such a case, the researcher inevitably has to seek permission from the care giver and the teacher.

The central principles in both situations are similar:

1 Demonstrate the importance of the study.

2 Deal with issues of confidentiality. This may involve assurance that the information will not be divulged to anyone and a demonstration of the safeguards that are in place to protect confidentiality. If a researcher intends to use quotations in their research report, it may be necessary to negotiate the procedures for doing this with the respondents. For example, do they want to be named or anonymous? Do they want to see a draft of the report?

3 Consent forms may sometimes be necessary. Even when you have received verbal consent from the respondent, a consent form may be required. An alternative is to use a negative consent letter that asks people to actively opt out of research. This can help reduce the number of people who do not return forms, but carries with it some ethical issues (such as the idea that people are automatically a part of the research unless they state otherwise).

Where researchers need to seek the permission of a third party such as a head teacher or parent, negotiating access can take some time and it is important to allow for this.

Some examples of applications for permission are provided in Appendix 2 at the end of this unit.

**Sample size**

What size of sample is required? Generally speaking, small samples are used to collect in-depth data and larger samples to collect survey-type data.

> In some real world research, the question [of sample size] is answered for you by the situation. You may be working in an organisation where the obvious thing is to survey everybody; or your resources may be so stretched that this sets the limits.
> (Robson, 2002: 161)

For fixed designs (survey or experiment) it is important to think about how the analysis is going to be conducted. The ‘rule of thumb’ is about 15 participants per variable in non-experimental designs (see also Robson, 2002). For surveys, Borg and Gall (cited in Robson) recommend about 100 observations for each of the major sub-groupings. Homogeneity within the population can make a difference. The less homogeneity, the larger the sample needed; this is true for both fixed and flexible designs.

Some phenomena can only show statistically significant effects when a large sample size is used. For example, there has been quite a lot of attention directed recently at girls' higher scores compared to those of boys in measures of reading skills and in national tests. However, these differences are quite small and will therefore only be statistically significant in large samples. This relationship between the size of the effect and the sample size, and indeed the amount of
variation in the population, is referred to as **statistical power** (see the discussion in Moser and Kalton (1971), Breakwell et al. (1995) and Robson (2002)). Where researchers use experiments it is important to undertake a **power analysis** (Clark-Carter (1997) provides power tables for this purpose - see Robson, 2002).

**Sampling on a small scale or for case studies**

Much educational research involves in-depth interviews and case studies. Here, the sample size may be very small and, because of this, researchers often attempt to ensure that a range of individuals or clusters are represented. So if you are sampling 20 primary schools, you may want to ensure that large cities, smaller towns and country schools are sampled (this is referred to as a heterogeneous sample). However, this is not always the approach that is taken. For example, in an in-depth study of 30 children's language experience at home Tizard and Hughes (1984) selected only white, English-born, working class girls for their study (a homogeneous sample). One would choose this approach to minimize the influence of extraneous factors on the focus of the investigation. Such decisions will, however, influence the extent to which you are justified in generalizing your findings.

The principles of generalizing from case studies or flexible designs are very different from those employed in surveys and experiments discussed above. When using flexible designs, findings can be generalized through theoretical proposition, either through proposing a new theory or testing the adequacy of an existing one to explain the case study data. Here the sampling strategies are sometimes referred to as **theoretical sampling** (also **purposive sampling**). This is in contrast to **statistical generalization**, which we have been dealing with up to now, where results are generalized to populations. In statistical generalization, an inference is made about a population on the basis of empirical data collected about a sample. This method of generalizing (statistically) is commonly recognized because research investigators have ready access to formulas for determining the confidence with which generalizations can be made, depending mostly upon the size and internal variation within the population and sample (used in surveys and experiments).

Where non-fixed designs are being used it can be more difficult to pre-specify the number of interviews or observations required. One formulation suggests that it is necessary to keep collecting data until ‘saturation’ is reached, in other words where the data are merely repeating or reinforcing what is already known and not producing new information or findings. Robson (2002: 199) talks about the factors that saturation depends on:

- the scope of the study - the broader the study, the longer it may take
- the nature of the topic - where there are obvious and clear topics a study may require fewer participants, but where a concern is more nebulous, more participants may be required
- study design - where there are a lot of data and analysis is done in great depth, fewer participants are needed.

**Depth versus breadth**

The trade-off between large but general studies and small, in-depth studies is between breadth and depth. If you want to gather information about the average amount of time children spend on reading in primary classrooms you will need to
use a largish sample. If you want detailed information about why some children spend more time on reading than others, and in what context, you will need an in-depth study. The limitation of survey data is that they may not give you detailed information on the meanings behind them. In contrast, it is hard (or impossible) to know the generalizability of in-depth data. The choice between these two levels of focus must be made in relation to the nature of the research questions being asked and the types of evidence required to address them.

Reliability and validity

Reliability

Synonyms for reliability include: dependability, stability, consistency, predictability, and accuracy (Kerlinger, 1986: Chapter 26). The main areas to consider in thinking about reliability in terms of questionnaires are **inter-rater reliability** and **test-retest reliability**.

Inter-rater reliability

If a data collection instrument is reliable it should not matter who is administering it (be it a reading test, a semi-structured questionnaire measuring coping skills, or other method); the application of the instrument should result in similar outcomes for the same subject (assuming nothing is different about the subject) whoever administers it - provided, of course, that they do so competently. For example, if two people test a child's reading using the same test under identical conditions they should arrive at very similar scores. If a tape recording of an interview is made, two raters should make similar assessments of the content of the interview. Often, the answers to open questions may be coded according to a pre-arranged coding strategy; for example, a question on a questionnaire to teachers about what they do when faced with a 'problem child' might be pre-coded in the following ways:

1 = deals with the problem  
2 = seeks someone else's help to deal{2}with the problem  
3 = avoids dealing with the problem.

In this sort of situation two raters should record the same rating for the same response. It is possible to check reliability by comparing the responses of two testers, interviewers, raters etc.

It is important to think about the likely effect of the interviewer on the respondent. The way the interviewer is dressed, their gender, ethnicity or social class, their position as expert, their age - all of these factors may have an effect on the interviewee, especially where the information being elicited is more personal or on a taboo subject. In some cases the setting may be important too. Is a child likely to tell a researcher their 'true' attitudes towards school if the teacher is present? Some information may be best gathered from a self-report, anonymous questionnaire. Contexts in which such sensitivity has proven to be problematic for face-to-face data gathering techniques include topics such as drug use and sexual activity (Davies, 1992).

How can you improve inter-rater reliability? This can be achieved by ensuring that the testers and raters use highly standardized procedures. In the case of
standardized tests this is achieved by giving careful and explicit instructions regarding administration and scoring. In the case of interviews the same applies, but where it is necessary to rate subjects who are responding in their own words it can be useful to give not only rating instructions but also ‘anchors’, or examples of a few responses for each code. Of course, not all interviews will be rated in this way and some data may be analysed using qualitative methods.

**Test-retest reliability**

Assuming that there is no reason to believe that someone has changed much between one measurement point and the next, they should score the same on a reading test, on an interview to measure their self-esteem etc. If they score very differently on the two occasions, the reliability of the instrument is said to be poor from the psychometric perspective.

**Internal consistency**

Rating scales and tests consist of a group of items that are put together to measure a particular construct reliably. Imagine a mathematics test: one question on its own will not enable an interviewer to measure a student's ability in maths and the researcher would need a range of questions to be able to test someone's ability effectively. The same is true of attitude scales and measures of self-esteem. The general rule for such tests is that the longer they are (the more test questions they contain), the more reliable they become. Having said that, there is of course a point beyond which reliability will not be increased by additional items and may decrease a researcher's response rate (Rust and Golombok, 1989).

The word ‘scale’ is used to describe the process of collating multiple measures of a given concept. De Vaus (1996) gives the following five reasons for using a scale:

1 It helps to focus on the complexity of a concept.

2 It assists in developing more valid measures. For example, one observation can be misleading or context-based, and additional items can minimize the significance of this sort of misinterpretation.

3 It increases reliability. An atypical response is cancelled out to some extent if there are further items.

4 It enables more precision. A single-question item does not enable us to differentiate very precisely between people.

5 Summarizing the answers to several questions as a scale total simplifies analysis - see how the Child Behaviour Questionnaire (Rutter, 1967) expresses the answers to a list of problems as one total score.

Scales with fewer than six items may not be reliable, but the individual item responses are rarely analysed and are instead summed to form a total score for the variable under investigation (like maths ability, level of self-esteem, and so on). The assumption is that the items are all measuring different facets of one construct; otherwise it would be incorrect to add them up to a total score. For example, you couldn't add up a student's performance on a maths test and a spelling test and then label the total as their maths ability. You therefore need to check that this assumption is correct and that the scale is internally consistent.
This also tends to pick up items that are confusingly worded or unreliable for other reasons.

The way to test for this is by calculating the statistic called 'Alpha', which measures how well the different items agree. Alpha ranges between 0 and 1, and is a function of the number of test items and the average inter-correlation among the items. Below, for conceptual purposes, the formula for the standardized Alpha is given:

\[
\alpha = \frac{N \bar{r}}{1 + (N-1) \bar{r}}
\]

Here \( N \) is equal to the number of items and \( \bar{r} \) (\( r \) with a bar above it) is the mean inter-item correlation among the items. You can see from this formula that if you increase the number of items, you increase the Alpha. Additionally, if the average inter-item correlation is low, Alpha will be low. As the average inter-item correlation increases, Alpha increases as well. The higher the figure for \( \alpha \), the more reliable is the scale. As a rule of thumb, Alpha should be at least 0.7 before it can be claimed that a scale is reliable (de Vaus, 1996). To increase Alpha, omit the unreliable items.

**Split-half reliability**

To calculate split-half reliability, a scale or test is randomly split into two to make two half-size versions of it. For each respondent, two scores are obtained, one for each half of the scale or test, and these are correlated with each other. The resultant correlation itself is not a reliability. We can obtain the reliability of the whole scale or test by applying what is called the 'Spearman-Brown' formula to this correlation:

\[
r_{test} = \frac{2r_{half}}{1 + r_{half}}
\]

where \( r_{test} \) is the reliability of the scale or test and \( r_{half} \) is the correlation obtained between the two halves of the instrument.

**Calculating reliability**

For statistical purposes, calculating the above reliabilities will require the use of what is called a correlation coefficient. Correlation coefficients need to be appropriate to the level of measurement or, in the case of nominal (categorical) data, the use of some measure of association. Kappa is the most frequently used in the latter case. So, the following statistical tests are the ones to use:

- interval data - Pearson's correlation
- ordinal data - Spearman's or Kendall's correlation
- nominal data - Cohen's Kappa.

There is no single number that is considered to be a cut-off for reliability. Oppenheim (1992) suggests that reliabilities below 0.8 are a problem. Rust and Golombok (1989) discuss the fact that reliabilities vary according to the subject matter of the test. IQ tests typically have reliabilities above 0.9 whereas for personality trait tests reliabilities of above 0.7 are expected. Essay marking tends
to produce notoriously low inter-rater reliabilities of about 0.6 and creativity tests (example: how many uses can you think of for a brick?) are about 0.5.

The significance of reliability

From the psychometric perspective, to be interpretable an instrument should be reliable. Unless you can depend upon the results of the measurement of your variables, you cannot with any confidence determine the relations between variables. The lack of an apparent relationship between variables may be due to the fact that they are not being measured well.

Reliability, while not the most important facet of measurement, is still extremely important. High reliability is no guarantee of good scientific results, but there can be no good scientific results without reliability. In brief, reliability is a necessary but not sufficient condition of the value of research results and their interpretation.

Validity

Essentially, questions of validity are concerned with whether or not an instrument measures what it purports to measure. For example, does a word-reading test measure reading ability or only some facet of that general construct of reading? Kerlinger (1986: 416) writes:

*The subject of validity is complex, controversial, and peculiarly important in behavioural research. Here perhaps more than anywhere else, the nature of reality is questioned.*

Typically, validity is subdivided into a range of categories (see Rust and Golombok (1989), Chapter 5).

Face validity

Face validity concerns the acceptability of the instrument items to both researcher and subject, for the operation being carried out. This should never be treated as trivial. If respondents fail to take the questions or task seriously, the results will be meaningless.

Content validity

Does the instrument cover all the aspects of the property being measured? For example, in the use of a selective test for employment, the content validity will be the extent to which the job specification matches the test specification. For many measures, content validation is basically judgemental. Usually, no one has defined all the possible aspects of the property being measured. Nonetheless, content validity is fundamental to psychometrics and other forms of measurement and is the main basis on which any test construction programme is judged. Content validity has to be judged qualitatively more often than quantitatively, as the form of any deviation from validity is usually more important than the degree (Rust and Golombok, 1989: 78).

Predictive validity

Predictive validity is used wherever tests are used to make predictions. How well does the measure being assessed affect subsequent performance in a related
area? For example, how well do school examinations at 14 years old predict job performance in the early 20s?

**Concurrent validity**

Concurrent validity describes the correlation between one measure and another that both purport to measure the same construct. Thus, a new intelligence test ought to correlate with existing intelligence tests.

**Construct validity**

This form of validity is the most fundamental of all. It applies to many forms of information gathering, and satisfactory construct validity implies the other forms of validity listed above. Does the measure satisfactorily represent the entity being investigated? The entity which the test is measuring is not normally measurable directly, and we are only really able to evaluate its usefulness by looking at the relationship between the test or measure and the various phenomena that the theory predicts. An example of construct validation is outlined by Rust and Golombok (1989). The Eysenck Personality Inventory measures extraversion/introversion and neuroticism. It was not possible for Eysenck to validate this scale by correlating respondents’ scores on an extroversion scale with their actual amount of extraversion. After all, if this were known there would be no need to have the scale. However, he was able, on the basis of his theory of extraversion, to suggest many ways in which extroverts might be expected to differ from introverts in their behaviour.

Construct validation is never complete, but is cumulative over a number of studies, and has many similarities to Popper's idea of verification in science (we cannot know what is true, but we can put forward refutable hypotheses on the basis of our conception of truth).

**Questionnaire piloting**

As has been already mentioned, designing your own questionnaire or rating scale is hard work. Some people devote years to devising questionnaires and checking their properties, reliability and validity. Most of us cannot afford to do this when working to tight deadlines but, however tight the deadline, piloting your questionnaire and documenting the piloting process are essential.

The piloting process, which is covered in detail in the readings for this unit from de Vaus (1996), essentially has three stages:

**Stage 1**

Question development - cognitive piloting (this process is declared to respondents)

Here you tell respondents that you are developing your questionnaire and ask them to comment on any issues they have with any of the items, wording, response categories, anything missing, etc. You then amend the questionnaire on the basis of the feedback.
**Stage 2**

Questionnaire development - dress rehearsal pilot (this process is undeclared to respondents)

Here researchers administer the questionnaire in circumstances that are as close as possible to those planned for the main study. The resulting data are analysed to determine whether the psychometric properties of the questionnaire are satisfactory. The central principles of questionnaire design are:

- **reliability**
- **validity**
- **discrimination**
  This refers to the idea that there may be a range of responses to a given question. Questions that everyone answers using the same categories do not discriminate between respondents and are therefore usually not very useful. If, for example, a researchers asks a question such as ‘Do you like living in a democracy?’ and 98% of your respondents answer ‘yes’, the question is largely uninformative.

- **response rate**
  If certain questions are systematically not completed, or answered with a ‘don’t know’, this suggests that the question is unclear or inappropriate.

- **same meaning for all respondents**
  This is mainly checked at the cognitive pilot stage.

- **relevance**
  In the process of analysing the pilot data, researchers need to think about whether this will really assist them in answering their research questions. Questionnaire length is always a problem - the shorter the better - so it is best not to ask irrelevant questions.

**Stage 3**

Polishing the pilot questionnaire Based on the feedback from the previous stage you can make final amendments to the questionnaire.