7

The nature of quantitative research

Chapter outline

Chapter guide

This chapter is concerned with the characteristics of quantitative research, an approach that has been the dominant strategy for conducting social research. Its influence has waned slightly since the mid-1970s, when qualitative research became increasingly influential. However, it continues to exert a powerful influence in many quarters. The emphasis in this chapter is very much on what quantitative research typically entails, though at a later point in the chapter the ways in which there are frequently departures from this ideal type are outlined. This chapter explores:

- the main steps of quantitative research, which are presented as a linear succession of stages;
- the importance of concepts in quantitative research and the ways in which measures may be devised for concepts; this discussion includes a discussion of the important idea of an *indicator*, which is devised as a way of measuring a concept for which there is no direct measure;
- the procedures for checking the reliability and validity of the measurement process;
- the main preoccupations of quantitative research, which are described in terms of four features: measurement; causality; generalization; and replication;
- some criticisms that are frequently levelled at quantitative research.

Introduction

In Chapter 2, quantitative research was outlined as a distinctive research strategy. In very broad terms, it was described as entailing the collection of numerical data, as exhibiting a view of the relationship between theory and research as deductive and a predilection for a natural science approach (and of positivism in particular), and as having an objectivist conception of social reality. A number of other features of quantitative research were outlined, but in this chapter we will be examining the strategy in much more detail.

It should be abundantly clear by now that the description of the research strategy as 'quantitative research' should not be taken to mean that quantification of aspects of social life is all that distinguishes it from a qualitative research strategy. The very fact that it has a distinctive epistemological and ontological position suggests that there is a good deal more to it than the mere presence of numbers. In this chapter, the main steps in quantitative research will be outlined. We will also examine some of the principal preoccupations of the strategy and how certain issues of concern among practitioners are addressed, such as the concerns about measurement validity.

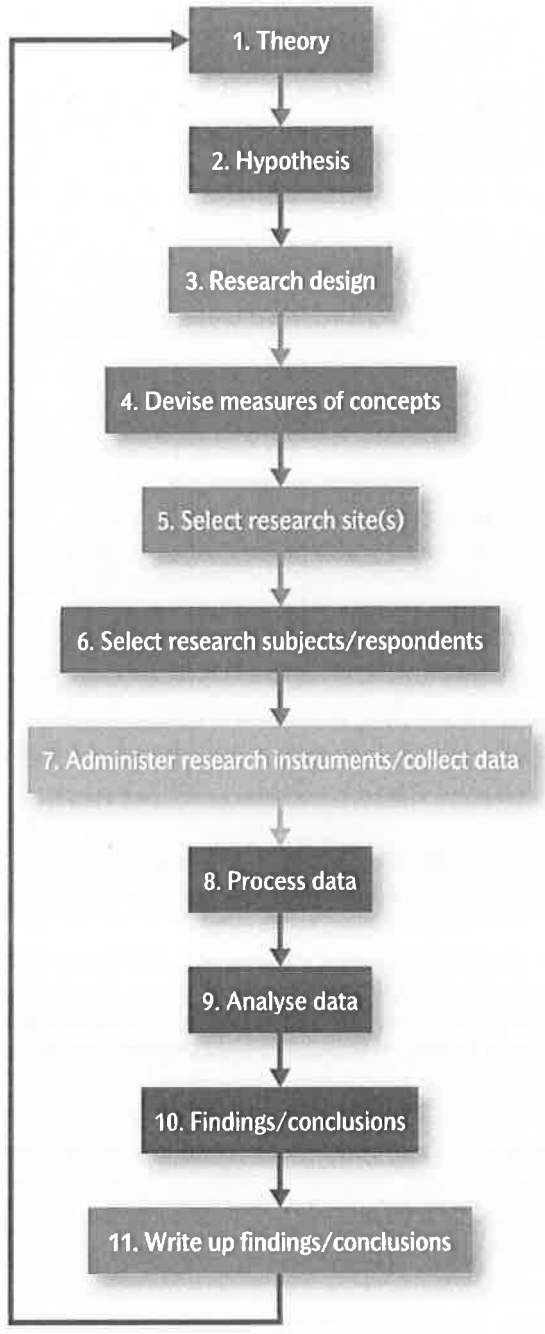
The main steps in quantitative research

Figure 7.1 outlines the main steps in quantitative research. This is very much an ideal-typical account of the process: it is probably never or rarely found in this pure form, but it represents a useful starting point for getting to grips with the main ingredients of the approach and the links between them. Research is rarely as linear and as straightforward as the figure implies, but its aim is to do no more than capture the main steps and to provide a rough indication of their interconnections.

Some of the chief steps have been covered in Chapters 1, 2, and 3. The fact that we start off with theory signifies that a broadly deductive approach to the relationship

Figure 7.1

The process of quantitative research



between theory and research is taken. It is common for outlines of the main steps of quantitative research to suggest that a hypothesis is deduced from the theory and is tested. This notion has been incorporated into Figure 7.1. However, a great deal of quantitative research does not entail the specification of a hypothesis, and instead theory acts loosely as a set of concerns in relation to which the social researcher collects data. The specification of hypotheses to be tested is particularly likely to be found in experimental research. Other research designs sometimes entail the testing of hypotheses. In Chapter 2 a study that employed a cross-sectional design using social survey research instruments was cited as an example (see Research in focus 2.4) that involved hypothesis testing. However, as a rule, we tend to find that Step 2 is more likely to be found in experimental research.

The next step entails the selection of a research design, a topic that was explored in Chapter 3. As we have seen, the selection of research design has implications for a variety of issues, such as the external validity of findings and researchers' ability to impute causality to their findings. Step 4 entails devising measures of the concepts in which the researcher is interested. This process is often referred to as *operationalization*, a term that originally derives from physics to refer to the operations by which a concept (such as temperature or velocity) is measured (Bridgman 1927). Aspects of this issue will be explored below in this chapter.

The next two steps entail the selection of a research site or sites and then the selection of subjects/respondents. (Experimental researchers tend to call the people on whom they conduct research 'subjects', whereas social survey researchers typically call them 'respondents'.) Thus, in social survey research an investigator must first be concerned to establish an appropriate setting for his or her research. A number of decisions may be involved. The well-known *Affluent Worker* research undertaken by Goldthorpe et al. (1968: 2–5) involved two decisions about a research site or setting. First, the researchers needed a community that would be appropriate for the testing of the 'embourgeoisement' thesis (the idea that affluent workers were becoming more middle class in their attitudes and lifestyles). As a result of this consideration, Luton was selected. Second, in order to come up with a sample of 'affluent workers' (Step 6), it was decided that people working for three of Luton's leading employers should be interviewed. Moreover, the researchers wanted the firms selected to cover a range of production technologies, because of evidence at that time that technologies had implications for workers' attitudes and behaviour. As a result of these considerations, the three firms were selected. Industrial workers were then sampled, again in terms of selected criteria that were to do with the researchers' interests in embourgeoisement and in the implications of technology for

work attitudes and behaviour. Research in focus 7.1 provides a more recent example of research that involved similar deliberations about selecting research sites and sampling respondents. In experimental research, these two steps are likely to include the assignment of subjects into control and treatment groups.

Research in focus 7.1
Selecting research sites and sampling respondents: the Social Change and Economic Life Initiative

The Social Change and Economic Life Initiative (SCELI) involved research in six labour markets: Aberdeen, Coventry, Kirkaldy, Northampton, Rochdale, and Swindon. These labour markets were chosen to reflect contrasting patterns of economic change in the early to mid-1980s and in the then recent past. Within each locality, three main surveys were carried out.

- *The Work Attitudes/Histories Survey*. Across the six localities a random sample of 6,111 individuals was interviewed using a structured interview schedule. Each interview comprised questions about the individual's work history and about a range of attitudes.
- *The Household and Community Survey*. A further survey was conducted on roughly one-third of those interviewed for the Work Attitudes/Histories Survey. Respondents and their partners were interviewed by structured interview schedule, and each person also completed a self-completion questionnaire. This survey was concerned with such areas as the domestic division of labour, leisure activities, and attitudes to the welfare state.
- *The Baseline Employers' Survey*. Each individual in each locality interviewed for the Work Attitudes/Histories Survey was asked to provide details of his or her employer (if appropriate). A sample of these employers was then interviewed by structured interview schedule. The interview schedules covered such areas as the gender distribution of jobs, the introduction of new technologies, and relationships with trade unions.

The bulk of the results was published in a series of volumes, including Penn et al. (1994) and A. M. Scott (1994). This example shows clearly the ways in which researchers are involved in decisions about selecting both research site(s) and respondents.

Step 7 involves the administration of the research instruments. In experimental research, this is likely to entail pre-testing subjects, manipulating the independent variable for the experimental group, and post-testing respondents. In cross-sectional research using social survey research instruments, it will involve interviewing the sample members by structured interview schedule or distributing a self-completion questionnaire. In research using structured observation, this step will mean an observer (or possibly more than one) watching the setting and the behaviour of people and then assigning categories to each element of behaviour.

Step 8 simply refers to the fact that, once information has been collected, it must be transformed into 'data'. In the context of quantitative research, this is likely to mean that it must be prepared so that it can be quantified. With some information this can be done in a relatively straightforward way—for example, for information relating to such things as people's ages, incomes, number of years spent at school, and so on. For other variables, quantification will entail *coding* the information—that is, transforming it into numbers to facilitate the quantitative analysis of the data, particularly if the analysis is going to be carried out by computer. Codes act as tags that are placed on data about people to allow the information to be processed by the computer. This consideration leads into Step 9—the analysis of the data. In this step, the researcher is concerned to use a number of techniques of quantitative data analysis to reduce the amount of data collected, to test for relationships between variables, to develop ways of presenting the results of the analysis to others, and so on.

On the basis of the analysis of the data, the researcher must interpret the results of the analysis. It is at this stage that the 'findings' will emerge. The researcher will consider the connections between the findings that emerge out of Step 8 and the various preoccupations that acted as the impetus of the research. If there is a hypothesis, is it supported? What are the implications of the findings for the theoretical ideas that formed the background to the research?

Then the research must be written up. It cannot take on significance beyond satisfying the researcher's personal curiosity until it enters the public domain in some way by being written up as a paper to be read at a conference or as a report to the agency that funded the research or as a book or journal article for academic social researchers. In writing up the findings and conclusions, the researcher is doing more than simply relaying what has been found to others: readers must be convinced that the research conclusions are important and that the findings are robust. Thus, a significant part of the research process entails convincing others of the significance and validity of one's findings.

Once the findings have been published, they become part of the stock of knowledge (or 'theory' in the loose sense of the word) in their domain. Thus, there is a feedback loop from Step 11 back up to Step 1. The presence of an element of both deductivism (Step 2) and inductivism (the feedback loop) is indicative of the positivist foundations of quantitative research. Similarly, the emphasis on the translation of concepts into measures (Step 4) is symptomatic of the principle of phenomenalism (see Key concept 2.2) that is also a feature of positivism. It is to this important phase of translating concepts into measures that we now turn. As we will see, certain considerations follow on from the stress placed on measurement in quantitative research. By and large, these considerations are to do with the validity and reliability of the measures devised by social scientists. These considerations will figure prominently in the following discussion.

As noted at the outset of presenting the model in Figure 7.1, this sequence of stages is a kind of ideal-typical account that is probably rarely found in this pure form. At the end of this chapter, the section 'Is it always like this?' deals with three ways in which the model may not be found in practice.

Concepts and their measurement

What is a concept?

Concepts are the building blocks of theory and represent the points around which social research is conducted. Just think of the numerous concepts that have already been mentioned in relation to research examples cited so far in this book:

> structure, agency, social class, job search method, deskilling, emotional satisfaction, religious orthodoxy, religious orientation, preservation of self, informal social control, negotiated order, culture, academic achievement, teacher expectations, charismatic leadership, healthy lifestyle, conversion.

Each represents a label that we give to elements of the social world that seem to have common features and that strike us as significant. As Bulmer (1984: 43) succinctly puts it, concepts 'are categories for the organisation of ideas and observations'. Thus, with a concept like social mobility, we notice that some people improve their socio-economic position relative to their parents, others stay roughly the same, and others are downwardly mobile. Out of such considerations, the concept of social mobility is reached.

If a concept is to be employed in quantitative research, it will have to be measured. Once they are measured, concepts can be in the form of independent or dependent variables. In other words, concepts may provide an explanation of a certain aspect of the social world, or they may stand for things we want to explain. A concept like social mobility may be used in either capacity: as a possible explanation of certain attitudes (are there differences between the downwardly mobile and others in terms of their political dispositions or social attitudes?) or as something to be explained (what are the causes of variation in social mobility?). Equally, we might be interested in evidence of changes in amounts of social mobility over time or in variations between comparable nations in levels of social mobility. As we start to investigate such issues, we are likely to formulate theories to help us understand why, for example, rates of social mobility vary between countries or over time. This will in turn generate new concepts, as we try to tackle the explanation of variation in rates.

Why measure?

There are three main reasons for the preoccupation with measurement in quantitative research.

1. Measurement allows us to delineate *fine differences* between people in terms of the characteristic in question. This is very useful, since, although we can often distinguish between people in terms of extreme categories, finer distinctions are much more difficult to recognize. We can detect clear variations in levels of job satisfaction—people who love their jobs and people who hate their jobs—but small differences are much more difficult to detect.
2. Measurement gives us a *consistent device* or yardstick for making such distinctions. A measurement device provides a consistent instrument for gauging differences. This consistency relates to two things: our ability to be consistent over time and our ability to be consistent with other researchers. In other words, a measure should be something that is influenced neither by the timing of its administration nor by the person who administers it. Obviously, saying that the measure is not influenced by timing is not meant to indicate that measurement readings do not change: they are bound to be influenced by the process of social change. What it means is that the measure should generate consistent results, other than those that occur as a result of natural changes. Whether a measure actually possesses this quality has to do with the issue of *reliability*, which was introduced in Chapter 3 and which will be examined again below.
3. Measurement provides the basis for *more precise estimates of the degree of relationship between concepts* (for example, through **correlation** analysis, which will be examined in Chapter 15). Thus, if we measure both job satisfaction and the things with which it might be related, such as stress-related illness, we will be able to produce more precise estimates of how closely they are related than if we had not proceeded in this way.

Indicators

In order to provide a measure of a concept (often referred to as an **operational definition**, a term deriving from the idea of operationalization), it is necessary to have an indicator or indicators that will stand for the concept (see Key concept 7.1). There are a number of ways in which indicators can be devised:

Key concept 7.1
What is an indicator?

It is worth making two distinctions here. First, there is a distinction between an *indicator* and a *measure*. The latter can be taken to refer to things that can be relatively unambiguously counted, such as personal income, household income, age, number of children, or number of years spent at school. Measures, in other words, are quantities. If we are interested in some of the causes of variation in personal income, the latter can be quantified in a reasonably direct way. We use indicators to tap concepts that are less directly quantifiable. If we are interested in the causes of variation in job satisfaction, we will need indicators that will stand for the concept. These indicators will allow job satisfaction to be measured, and we can treat the resulting quantitative information as if it were a measure. An indicator, then, is something that is devised or already exists and that is employed *as though it were a measure of a concept*. It is viewed as an indirect measure of a concept, like job satisfaction. We see here a second distinction between *direct* and *indirect* indicators of concepts. Indicators may be direct or indirect in their relationship to the concepts for which they stand. Thus, an indicator of marital status has a much more direct relationship to its concept than an indicator (or set of indicators) relating to job satisfaction. Sets of attitudes always need to be measured by batteries of indirect indicators. So too do many forms of behaviour. When indicators are used that are not true quantities, they will need to be coded to be turned into quantities. Directness and indirectness are not qualities inherent to an indicator: data from a survey question on amount earned per month may be a direct measure of personal income. However, if we treat personal income as an indicator of social class, it becomes an indirect measure. The issue of indirectness raises the question of where an indirect measure comes from—that is, how does a researcher devise an indicator of something like job satisfaction? Usually, it is based on common-sense understandings of the forms the concept takes or on anecdotal or qualitative evidence relating to that concept.

- through a question (or series of questions) that is part of a structured interview schedule or self-completion questionnaire; the question(s) could be concerned with the respondents' report of an attitude (for example, job satisfaction) or their social situation (for example, poverty) or a report of their behaviour (for example, leisure pursuits);
- through the recording of individuals' behaviour using a structured observation schedule (for example, pupil behaviour in a classroom);
- through official statistics, such as the use of Home Office crime statistics to measure criminal behaviour;
- through an examination of mass media content through content analysis—for example, to determine changes in the salience of an issue, such as AIDS, in the mass media (Beharrell 1993).

Indicators, then, can be derived from a wide variety of different sources and methods. Very often the researcher has to consider whether one indicator of a concept will be sufficient. This consideration is frequently a focus for social survey researchers. Rather than have just a single indicator of a concept, the researcher may feel that it may be preferable to ask a number of questions in the course of a structured interview or a self-completion questionnaire that tap a certain concept (see Research in focus 7.2 and 7.3 for examples).

Research in focus 7.2
A multiple-indicator measure of a concept

The research on the effects of redundancy by Westergaard et al. (1989), which was referred to in Chapters 2 and 3, was conducted by structured interview with 378 steel workers who had been made redundant. One of the authors' interests was whether their respondents' commitment to work varied according to whether they were still unemployed at the time of the interview or had found work or had retired. In order to measure commitment to employment, the authors gave their respondents ten statements and asked them to indicate their level of agreement or disagreement on a seven-point scale running from 'Yes, I strongly agree' to 'No, I strongly disagree'. There was a middle point on the scale that allowed for a neutral response. This approach to investigating a cluster of attitudes is known as a **Likert scale**, though in many cases researchers use a five-point rather than a seven-point scale for responses. See Key concept 7.2 for a description of what a Likert scale entails. The ten statements were as follows.

1. Work is necessary, but rarely enjoyable.
2. Having a job is not very important to me.
3. I regard time spent at work as time taken away from the things I want to do.
4. Having a job is/was important to me only because it brings in money.
5. Even if I won a great deal of money on the pools I'd carry on working.
6. If unemployment benefit were really high, I would still prefer to work.
7. I would hate being on the dole.
8. I would soon get bored if I did not go out to work.
9. The most important things that have happened to me involved work.
10. Any feelings I've had in the past of achieving something worthwhile have usually come through things I've done at work.

In fact, the authors found that their respondents' replies did not differ a great deal in terms of whether they had found work since being made redundant or were still unemployed or had taken retirement.

Key concept 7.2
What is a Likert scale?

The investigation of attitudes is a prominent area in much survey research. One of the most common techniques for conducting such an investigation is the Likert scale, named after Rensis Likert, who developed the method. The Likert scale is essentially a **multiple-indicator** or **multiple-item measure** of a set of attitudes relating to a particular area. The goal of the Likert scale is to measure intensity of feelings about the area in question. In its most common format, it comprises a series of statements (known as 'items') that focus on a certain issue or theme. Each respondent is then asked to indicate his or her level of agreement with the statement. Usually, the format for indicating level of agreement is a five-point scale going from 'strongly agree' to 'strongly disagree', but seven-point scale and other formats are used too. There is usually a middle position of 'neither agree nor disagree' or 'undecided' indicating neutrality on the issue. Each respondent's reply on each item is scored, and then the scores for each item are aggregated to form an overall score. Normally, since the scale measures intensity, the scoring is carried out so that a high level of intensity of feelings in connection with each indicator receives a high score (for example, on a five-point scale, a score of 5 for very strong positive feelings about an issue and a score of 1 for very negative feelings). The measure of commitment to work referred to in Research in focus 7.2 is an example of a Likert scale. Variations on the typical format of indicating degrees of agreement are scales referring to frequency (for example, 'never' through to 'always') and evaluation (for example, 'very poor' through to 'very good').

There are several points to bear in mind about the construction of a Likert scale. The following are particularly important.

- The items must be statements and not questions.
- The items must all relate to the same object (job, organization, ethnic groups, unemployment, sentencing of offenders, etc.).
- The items that make up the scale should be interrelated (see the discussion of **internal reliability** in this chapter and Key concept 7.3).

It is useful to vary the phrasing so that some items imply a positive view of the phenomenon of interest and others a negative one. Thus, in the example in Research in focus 7.2, some items imply a negative view of work (for example, 'Having a job is not very important to me') and others a positive view of work (for example, 'I would soon get bored if I did not go out to work'). This variation is advised in order to identify respondents who exhibit **response sets** (see the sections on 'Response sets' in Chapters 9 and 10).

Using multiple-indicator measures

What are the advantages of using a multiple-indicator measure of a concept? The main reason for their use is a recognition that there are potential problems with a reliance on just a single indicator:

- It is possible that a single indicator will incorrectly classify many individuals. This may be due to the wording of the question or it may be a product of misunderstanding. But, if there are a number of indicators, if people are misclassified through a particular question, it will be possible to offset its effects.
- One indicator may capture only a portion of the underlying concept or be too general. A single question may need to be of an excessively high level of generality and so may not reflect the true state of affairs for the people replying to it. Alternatively, a question may cover only one aspect of the concept in question. For example, if you were interested in job satisfaction, would it be sufficient to ask people how satisfied they were with their pay? Almost certainly not, because most people would argue that there is more to job satisfaction than just satisfaction with pay. A single indicator such as this would be missing out on such things as satisfaction with conditions, with the work itself, and with other aspects of the work environment. By asking a number of questions, the researcher can get access to a wider range of aspects of the concept.

- You can make much finer distinctions. Taking the Westergaard et al. (1989) measure of commitment to work as an example (see Research in focus 7.2), if we just took one of the indicators as a measure, we would be able to array people only on a scale of 1 to 7, assuming that answers indicating no commitment were assigned 1 and answers indicating a very high level of commitment were assigned 7, the five other points being scored 2, 3, 4, 5, and 6. However, with a multiple-indicator measure of ten indicators the range is 10 (10 × 1) – 70 (10 × 7). Key concept 7.2 provides some information about the kind of scale (a Likert scale) that was used in the study by Westergaard et al.

Research in focus 7.3
A multiple-indicator measure of another concept

In Kelley and De Graaf's (1997) research on religious beliefs, two of the main concepts in which they were interested—national religiosity and family religious orientation—were each measured by a single indicator (see Research in focus 2.4). However, religious orthodoxy was measured by four survey questions, answers to which were aggregated for each respondent to form a 'score' for that person. Answers to each of the four questions were given a score and then aggregated to form a religious belief score. The four questions were as follows.

1. Please indicate which statement below comes closest to expressing what you believe about God:
   - I don't believe in God.
   - I don't know whether there is a God and I don't believe there is any way to find out.
   - I don't believe in a personal God, but I do believe in a higher power of some kind.
   - I find myself believing in God some of the time, but not at others.
   - While I have doubts, I feel that I do believe in God.
   - I know God really exists and I have no doubts about it.
2. Which best describes your beliefs about God?
   - I don't believe in God and I never have.
   - I don't believe in God, but I used to.
   - I believe in God now, but I didn't used to.
   - I believe in God now and I always have.
3. How close do you feel to God most of the time?
   - Don't believe in God.
   - Not close at all.
   - Not very close.
   - Somewhat close.
   - Extremely close.
4. There is a God who concerns Himself with every human being, personally.
   - Strongly agree.
   - Agree.
   - Neither agree nor disagree.
   - Disagree.
   - Strongly disagree.

Dimensions of concepts

One elaboration of the general approach to measurement is to consider the possibility that the concept in which you are interested comprises different **dimensions**. This view is particularly associated with Lazarsfeld (1958). The idea behind this approach is that, when the researcher is seeking to develop a measure of a concept, the

different aspects or components of that concept should be considered. This specification of the dimensions of a concept would be undertaken with reference to theory and research associated with that concept. Examples of this kind of approach can be discerned in Seeman's (1959) delineation of five dimensions of alienation (powerlessness, meaninglessness, normlessness, isolation, and self-estrangement). Bryman and Cramer (2011) demonstrate the operation of this approach with reference to the concept of 'professionalism'. The idea is that people scoring high on one dimension may not necessarily score high on other dimensions, so that for each respondent you end up with a multidimensional 'profile'. Research in focus 7.4 demonstrates the use of dimensions in connection with the concept of 'deskilling' in the sociology of work.

However, in much if not most quantitative research, there is a tendency to rely on a single indicator of concepts. For many purposes this is quite adequate. It would be a mistake to believe that investigations that use a single indicator of core concepts are somehow deficient. In any case, some studies, like Kelley and De Graaf (1997, see Research in focus 7.3), employ both single- and multiple-indicator measures of concepts. What *is* crucial is whether measures are reliable and whether they are valid representations of the concepts they are supposed to be tapping. It is to this issue that we now turn.

Research in focus 7.4
Specifying dimensions of a concept: the case of deskilling

This example is taken from social survey research primarily concerned with social class in Britain by Marshall et al. (1988). The research was based on structured interviews with a national, random sample of individuals. One of the researchers' areas of interest was Braverman's (1974) deskilling thesis (see Research in focus 2.2). Based on a reading of the literature on this topic at the time, the authors argued that two important components or *dimensions* of deskilling on which they were able to shed light were 'skill as complexity and skill as freedom', which 'are central to the thesis that work is being proletarianized through the deskilling of tasks' (Marshall et al. 1988: 116). 'Skill as complexity' was measured by a single interview question asking respondents whether their current jobs required more, less, or about the same amount of skill as when they first started. 'Skill as freedom' was measured by seven indicators that were treated separately and not aggregated. The questions entailed asking respondents about such things as whether they were able to reduce the pace of their work or to initiate new tasks in their work. Neither dimension comprised measures that offered significant support for the deskilling thesis.

Reliability and validity

Although the terms 'reliability' and 'validity' seem to be almost synonymous, they have quite different meanings in relation to the evaluation of measures of concepts, as was seen in Chapter 3.

Reliability

As Key concept 7.3 suggests, reliability is fundamentally concerned with issues of consistency of measures. There are at least three different meanings of the term. These are outlined in Key concept 7.3 and elaborated upon below.

Stability

The most obvious way of testing for the stability of a measure is the *test–retest* method. This involves administering a test or measure on one occasion and then readministering it to the same sample on another occasion—that is:

T_1	T_2
Obs_1	Obs_2

We should expect to find a high correlation between Obs_1 and Obs_2. Correlation is a measure of the strength of the relationship between two variables. This topic will be covered in Chapter 15 in the context of a discussion about quantitative data analysis. Let us imagine that we develop a multiple-indicator measure that is supposed to tap a concept that we might call 'designerism' (a preference for buying goods and especially clothing with 'designer' labels). We would administer the measure to a sample of respondents and readminister it some time later. If the correlation is low, the measure would appear to be unstable, implying that respondents' answers cannot be relied upon.

Key concept 7.3
What is reliability?

Reliability refers to the consistency of a measure of a concept. The following are three prominent factors involved when considering whether a measure is reliable:

- *Stability*. This consideration entails asking whether a measure is stable over time, so that we can be confident that the results relating to that measure for a sample of respondents do not fluctuate. This means that, if we administer a measure to a group and then readminister it, there will be little variation over time in the results obtained. In February 2010, the then Shadow Home Secretary, Chris Grayling, was roundly criticized by the UK Statistics Authority for comparing Home Office statistics from the late 1990s with current figures to suggest that there had been a big increase in violent crimes since Labour took office in 1997. The reason for the criticism was that there had been a change to the definition of violent crime, which had produced an immediate 35 per cent increase in the crime. In this case, the measure of violent crime was not reliable from the point of view of inferring a change over time. For this story, see 'Chris Grayling Accused of Damaging Public Trust over Crime Figures', **www.thetimes.co.uk/tto/news/politics/article2030815.ece** (accessed 9 August 2010).
- *Internal reliability*. The key issue is whether the indicators that make up the **scale** or **index** are consistent—in other words, whether respondents' scores on any one indicator tend to be related to their scores on the other indicators.
- *Inter-observer consistency*. When a great deal of subjective judgement is involved in such activities as the recording of observations or the translation of data into categories and where more than one 'observer' is involved in such activities, there is the possibility that there is a lack of consistency in their decisions. This can arise in a number of contexts, for example: in content analysis where decisions have to be made about how to categorize media items; when answers to open questions have to be categorized; or in structured observation when observers have to decide how to classify subjects' behaviour.

However, there are a number of problems with this approach to evaluating reliability. Respondents' answers at T_1 may influence how they reply at T_2. This may result in greater consistency between Obs_1 and Obs_2 than is in fact the case. Second, events may intervene between T_1 and T_2 that influence the degree of consistency. For example, if a long span of time is involved, changes in the economy or in respondents' personal financial circumstances could influence their views about and predilection for designer goods. For example, Berthoud (2000b) notes that an index of ill-health devised from the British Household Panel Survey (BHPS) achieved a high test–retest reliability. He notes that this is very encouraging, because 'some of the variation between tests (a year apart) will have been caused by genuine changes in people's health' (Berthoud 2000b: 170). There is no easy way of disentangling the effects of a lack of stability in the measure from 'real' changes in people's health over the year in question.

There are no clear solutions to these problems, other than by introducing a complex research design and so turning the investigation of reliability into a major project in its own right. Perhaps for these reasons, many

if not most reports of research findings do not appear to carry out tests of stability. Indeed, longitudinal research is often undertaken precisely in order to identify social change and its correlates.

Internal reliability

This meaning of reliability applies to multiple-indicator measures like those examined in Research in focus 7.2 and 7.3. When you have a multiple-item measure in which each respondent's answers to each question are aggregated to form an overall score, the possibility is raised that the indicators do not relate to the same thing; in other words, they lack coherence. We need to be sure that all our designerism indicators are related to each other. If they are not, some of the items may actually be unrelated to designerism and therefore indicative of something else.

One way of testing internal reliability is the *split-half* method. We can take the commitment to work measure developed by Westergaard et al. (1989) as an example (see Research in focus 7.2). The ten indicators would be divided into two halves with five in each group. The indicators would be allocated on a random or an odd–even basis. The degree of correlation between scores on two halves would then be calculated. In other words, the aim would be to establish whether respondents scoring high on one of the two groups also scored high on the other group of indicators. The calculation of the correlation will yield a figure, known as a coefficient, that varies between 0 (no correlation and therefore no internal consistency) to 1 (perfect correlation and therefore complete internal consistency). It is usually expected that a result of 0.80 and above implies an acceptable level of internal reliability. Do not worry if the figures appear somewhat opaque. The meaning of correlation will be explored in much greater detail later on. The chief point to carry away with you at this stage is that the correlation establishes how closely respondents' scores on the two groups of indicators are related.

Nowadays, most researchers use a test of internal reliability known as *Cronbach's alpha* (see Key concept 7.4). Its use has grown as a result of its incorporation into computer software for quantitative data analysis.

Inter-observer consistency

The idea of inter-observer consistency is briefly outlined in Key concept 7.3. The issues involved are rather too advanced to be dealt with at this stage and will be touched on briefly in later chapters. Cramer (1998: ch. 14) provides a very detailed treatment of the issues and appropriate techniques.

Key concept 7.4
What is Cronbach's alpha?

To a very large extent we are leaping ahead too much here, but it is important to appreciate the basic features of what this widely used test means. Cronbach's alpha is a commonly used test of internal reliability. It essentially calculates the average of all possible split-half reliability coefficients. A computed alpha coefficient will vary between 1 (denoting perfect internal reliability) and 0 (denoting no internal reliability). The figure 0.80 is typically employed as a rule of thumb to denote an acceptable level of internal reliability, though many writers work with a slightly lower figure. In the case of the commitment to work scale devised by Westergaard et al. (1989: 93), alpha was 0.70, which they refer to as 'a satisfactory level'. In the case of Kelley and De Graaf's (1997) measure of religious orthodoxy, which comprised four indicators, alpha was 0.93. The alpha levels varied between 0.79 and 0.95 for each of the fifteen national samples that make up the data. Berthoud (2000b: 169) writes that a minimum level of 0.60 is 'good' and cites the case of an index of ill-health used in the BHPS that achieved a level of 0.77.

Validity

As noted in Chapter 3, the issue of measurement validity has to do with whether a measure of a concept really measures that concept (see Key concept 7.5). When people argue about whether a person's IQ score really measures or reflects that person's level of intelligence, they are raising questions about the measurement validity of the IQ test in relation to the concept of intelligence. Similarly, one often hears people say that they do not believe that the Retail Price Index really reflects inflation and the rise in the cost of living. Again, a query is being raised in such comments about measurement validity. And whenever students or lecturers debate whether

formal examinations provide an accurate measure of academic ability, they too are raising questions about measurement validity.

Writers on measurement validity distinguish between a number of ways of appraising measurement validity. These types really reflect different ways of gauging the validity of a measure of a concept. These different ways of appraising measurement validity will now be outlined.

Key concept 7.5
What is validity?

Validity refers to the issue of whether an indicator (or set of indicators) that is devised to gauge a concept really measures that concept. Several ways of establishing validity are explored in the text: face validity; concurrent validity; predictive validity; construct validity; and convergent validity. Here the term is being used as a shorthand for what was referred to as *measurement validity* in Chapter 3. Validity should therefore be distinguished from the other terms introduced in Chapter 3: internal validity; external validity; and ecological validity.

Face validity

At the very minimum, a researcher who develops a new measure should establish that it has **face validity**—that is, that the measure apparently reflects the content of the concept in question. Face validity might be established by asking other people whether the measure seems to be getting at the concept that is the focus of attention. In other words, people, possibly those with experience or expertise in a field, might be asked to act as judges to determine whether on the face of it the measure seems to reflect the concept concerned. Face validity is, therefore, an essentially intuitive process.

Concurrent validity

The researcher might seek also to gauge the **concurrent validity** of the measure. Here the researcher employs a *criterion* on which cases (for example, people) are known to differ and that is relevant to the concept in question. A new measure of job satisfaction can serve as an example. A criterion might be absenteeism, because some people are more often absent from work (other than through illness) than others. In order to establish the concurrent validity of a measure of job satisfaction, we might see how far people who are satisfied with their jobs are less likely than those who are not satisfied to be *absent* from work. If a lack of correspondence were found, such as there being no difference in levels of job satisfaction among frequent absentees, doubt might be cast on whether our measure is really addressing job satisfaction. Similarly, Wood and Williams (2007) discuss the problem of asking people in questionnaires how much they spend on gambling, because self-reported gambling expenditure tends to be inconsistent with actual revenue that accrues from gambling. The authors asked a large random sample of residents in Ontario, Canada, how much they had spent in the last month in twelve different ways. They note that even slight variations in the wording of questions could result in very different estimates of expenditure on the part of respondents, a concern that relates to issues that are discussed in Chapter 11. However, some questions did produce answers that were more consistent with an estimate of gambling expenditure per person in Ontario, which acted as the concurrent validity criterion. The authors recommend on the basis of its performance in the validity test and its face validity the following question:

> Roughly how much money do you spend on [specific gambling activity] in a typical month? What we mean here is how much you are ahead or behind, or your net win or loss in a typical month. (Wood and Williams 2007: 68)

The question required aggregating respondents' estimates of their gambling expenditure on each of several gambling activities.

Predictive validity

Another possible test for the validity of a new measure is *predictive validity*, whereby the researcher uses a *future* criterion measure, rather than a contemporary one, as in the case of concurrent validity. With predictive validity, the researcher would take future levels of absenteeism as the criterion against which the validity of a new measure

of job satisfaction would be examined. The difference from concurrent validity is that a future rather than a simultaneous criterion measure is employed.

Construct validity

Some writers advocate that the researcher should also estimate the *construct validity* of a measure. Here, the researcher is encouraged to deduce hypotheses from a theory that is relevant to the concept. For example, drawing upon ideas about the impact of technology on the experience of work, the researcher might anticipate that people who are satisfied with their jobs are less likely to work on routine jobs; those who are not satisfied are more likely to work on routine jobs. Accordingly, we could investigate this theoretical deduction by examining the relationship between job satisfaction and job routine. However, some caution is required in interpreting the absence of a relationship between job satisfaction and job routine in this example. First, either the theory or the deduction that is made from it might be misguided. Second, the measure of job routine could be an invalid measure of that concept.

Convergent validity

In the view of some methodologists, the validity of a measure ought to be gauged by comparing it to measures of the same concept developed through other methods. For example, if we develop a questionnaire measure of how much time managers spend on various activities (such as attending meetings, touring their organization, informal discussions, and so on), we might examine its validity by tracking a number of managers and using a structured observation schedule to record how much time is spent in various activities and their frequency. In addition to using a test of concurrent validity for their research on gambling expenditure, Wood and Williams (2007) used a diary to estimate gambling expenditure for a subsample of their respondents that could then be compared to questionnaire estimates. Respondents began the diary shortly after they had answered the survey question and continued completing it for a thirty-day period. This validity test allowed the researchers to compare what was actually spent in the month after the question was asked (assuming the diary estimates were correct) with what respondents *thought* they spent on gambling.

An interesting instance of convergent *in*validity is described in Thinking deeply 7.1. In this example, the British Crime Survey (BCS) was consciously devised to provide an alternative measure of levels of crime so that it would act as a check on the official statistics. The two sets of data are collected in quite different ways: the official crime statistics are collected as part of the

Thinking deeply 7.1
A case of convergent *in*validity: Home Office crime statistics

An article in the *Sunday Times* (Burrell and Leppard 1994) proclaimed the government's claims about the fall in crime a sham. The opening paragraph put the point as follows:

> The government's much heralded fall in crime is a myth. Hundreds of thousands of serious crimes have been quietly dropped from police records as senior officers massage their statistics to meet new Home Office targets. . . . Crime experts say at least 220,000 crimes, including burglary, assault, theft and car crimes, vanished from official statistics last year as a result of police manipulation of the figures.

What gave the 'crime experts' and the reporters the confidence to assert that the much-trumpeted fall in crime was a myth because the figures on which the claim was made had been massaged? The answer is that data from the British Crime Survey (BCS) had 'recently reported that actual crime rose faster over the past two years than during the 1980s' (see Research in focus 7.2 for details of the BCS). In each case, a large, randomly selected sample of individuals is questioned by structured interview. The survey is not based on a panel research design, because the same people are not interviewed with each wave of data collection. The BCS is an example of what is known as a 'victimization survey'. With this kind of survey, a sample of a population is questioned about its experiences as victims of crime. The idea is that unreported crime and other crime that does not show up in the official statistics will be revealed. The categories of crime used in the survey are meant to reflect those reported in the official statistics (Coleman and Moynihan 1996: 83–6). The 1994 survey found that there had been a marked increase in most categories of crime.

bureaucratic processing of offenders in the course of the activities of members of the British criminal justice system, whereas the BCS entails the collection of data by interview from a national sample of possible victims of crime. In the case reported in Thinking deeply 7.1 a lack of convergent validity was found. However, the problem with the convergent approach to testing validity is that it is not possible to establish very easily which of the two measures represents the more accurate picture. The BCS is not entirely flawless in its approach to the measurement of crime levels, and, in any case, the 'true' picture with regard to the volume of crime at any one time is an almost entirely metaphysical notion (Reiner 2000b). While the authors of the news item were able to draw on bits of anecdotal evidence to support their thesis that the figures were being massaged and this together with the BCS evidence casts doubt on the official statistics, it would be a mistake to hold that the survey evidence necessarily represents a definitive and therefore unambiguously valid measure.

Research in focus 7.5 provides a brief account of a new scale using the Likert procedure and some of the ways in which reliability and validity were assessed.

Reflections on reliability and validity

There are, then, a number of different ways of investigating the merit of measures that are devised to represent social scientific concepts. However, the discussion of reliability and validity is potentially misleading, because it would be wrong to think that all new measures of concepts are submitted to the rigours described above. In fact, most typically, measurement is undertaken within a stance that Cicourel (1964) described as 'measurement by fiat'. By the term 'fiat', Cicourel was referring not to a well-known Italian car manufacturer but to the notion of 'decree'. He meant that most measures are simply asserted. Fairly straightforward but minimal steps may be taken to ensure that a measure is reliable and/or valid, such as testing for internal reliability when a multiple-indicator measure has been devised and examining face validity. But in many if not the majority of cases in which a concept is measured, no further testing takes place. This point will be further elaborated below.

It should also be borne in mind that, although reliability and validity are analytically distinguishable, they are related because validity presumes reliability. This means that, if your measure is not reliable, it cannot be valid (see page 47). This point can be made with respect to each of the three criteria of reliability that have been discussed. If the measure is not stable over time, it simply cannot be providing a valid measure. The measure could not be tapping the concept it is supposed to be related to if the measure fluctuated. If the measure fluctuates, it may be measuring different things on different occasions. If a measure lacks internal reliability, it means that a multiple-indicator measure is actually measuring two or more different things. Therefore, the measure cannot be valid. Finally, if there is a lack of inter-observer consistency, it means that observers cannot agree on the meaning of what they are observing, which in turn means that a valid measure cannot be in operation.

Research in focus 7.5
Developing a Likert scale: the case of attitudes to vegetarians

Chin et al. (2002) describe how they went about developing a scale designed to measure pro- or anti-vegetarian attitudes. They note that non-vegetarians sometimes see vegetarianism as deviant and that, as a result, vegetarians are sometimes regarded with suspicion if not hostility. The authors developed a scale comprising thirty-three items. Each item is a statement to which the respondent is asked to indicate strength of agreement or disagreement on a seven-point scale. The items were arrived at following: interviews with both vegetarians and non-vegetarians; a review of the literature on vegetarianism; field observations (though it is not clear of what or whom); brainstorming within the team; and an examination of attitude scales addressing other forms of prejudice for possible wording and presentation. The items were meant to tap four areas:

- forms of behaviour in which vegetarians engage that are viewed as irritating—for example, 'Vegetarians preach too much about their beliefs and eating habits' (possibly a double-barrelled item—see Chapter 11);
- disagreement with vegetarians' beliefs—for example, 'Vegetarians are overly concerned with animal rights';

- health-related aspects of being a vegetarian—for example, 'Vegetarians are overly concerned about gaining weight';
- appropriate treatment of vegetarians—for example, 'It's OK to tease someone for being a vegetarian'.

The scale was tested out on a sample of university undergraduates in the USA. Some items from the scale were dropped because they exhibited poor internal consistency with the other items. Cronbach's alpha was conducted for the remaining twenty-one items and found to be high at 0.87 (see Key concept 7.4). The construct validity (see above on the meaning of this term) of the scale was also tested by asking the students to complete other scales that the researchers predicted would be associated with pro- or anti-vegetarian attitudes. One method was that the authors hypothesized that people with authoritarian attitudes would be more likely to be anti-vegetarians. This was confirmed, although the relationship between these two variables was very small. However, contrary to their hypothesis, the scale for attitudes towards vegetarianism was *not* found to be related to political conservatism. The scale emerges as internally reliable (see Key concept 7.3 on the meaning of this term) but as being of slightly questionable construct validity.

Research in focus 7.6
Assessing the internal reliability and the concurrent and predictive validity of a measure of organizational climate

Patterson et al. (2005) describe the way they went about validating a measure they developed of organizational climate. This is a rather loose concept that was first developed in the 1960s and 1970s to refer to the perceptions of an organization by its members. Four main dimensions of climate were developed based around the following notions:

1. *human relations model*: feelings of belonging and trust in the organization and the degree to which there is training, good communication, and supervisory support;
2. *internal process model*: the degree of emphasis on formal rules and on traditional ways of doing things;
3. *open systems model*: the extent to which flexibility and innovativeness are valued;
4. *rational goal model*: the degree to which clearly defined objectives and the norms and values associated with efficiency, quality, and high performance are emphasized.

An Organizational Climate Measure, comprising 95 items in a four-point Likert format (definitely false, mostly false, mostly true, definitely true) was developed and administered to employees in 55 UK organizations, with 6,869 completing a questionnaire—a response rate of 57 per cent. A **factor analysis** (see Key concept 7.6) was conducted to explore the extent to which there were distinct groupings of items that tended to go together. This procedure yielded seventeen scales, such as autonomy, involvement, innovation and flexibility, and clarity of organizational goals.

The *internal reliability* of the scales was assessed using Cronbach's alpha, showing that all scales were at a level of 0.73 or above. This suggests that the measure's constituent scales were internally reliable.

Concurrent validity was assessed following semi-structured interviews, with each company's managers in connection with their organization's practices. The interview data were coded to provide criteria against which the validity of the scales could be gauged. In most cases, the scales were found to be concurrently valid. For example, the researcher examined the correlation between a scale designed to measure the emphasis on tradition and the degree to which practices associated with the 'new manufacturing paradigm' (Patterson et al. 2005: 397) were adopted, as revealed by the interview data. The correlation was –0.42, implying that those firms

that were perceived as rooted in tradition tended to be less likely to adopt new manufacturing practices. Here the adoption of new manufacturing practices was treated as a criterion to assess the extent to which the scale measuring perceptions of tradition really was addressing tradition. If the correlation had been small or had been positive, the concurrent validity of the scale would have been in doubt.

To assess *predictive validity*, the researchers asked a senior **key informant** at each company to complete a questionnaire one year after the main survey had been conducted. The questionnaire was meant to address two of the measure's constituent scales, one of which was the innovation and flexibility scale. It asked the informants to assess their company in terms of its innovativeness in a number of areas. For example, the correlation between the innovation and flexibility scale and informants' assessments of their companies in terms of innovativeness with respect to products achieved a correlation of 0.53. This implies that there was indeed a correlation between perceptions of innovativeness and flexibility and a subsequent indicator of innovativeness.

The main preoccupations of quantitative researchers

Both quantitative and qualitative research can be viewed as exhibiting a set of distinctive but contrasting preoccupations. These preoccupations reflect epistemologically grounded beliefs about what constitutes acceptable knowledge. In this section, four distinctive preoccupations that can be discerned in quantitative research will be outlined and examined: measurement, causality, generalization, and replication.

Measurement

The most obvious preoccupation is with measurement, a feature that is scarcely surprising in the light of much of the discussion in the present chapter so far. From the position of quantitative research, measurement carries a number of advantages that were previously outlined. It is not surprising, therefore, that issues of reliability and validity are a concern for quantitative researchers, though this is not always manifested in research practice.

Causality

There is a very strong concern in most quantitative research with explanation. Quantitative researchers are rarely concerned merely to describe how things are, but are keen to say why things are the way they are. This emphasis is also often taken to be a feature of the ways in which the natural sciences proceed. Thus, researchers are often not only interested in a phenomenon like racial prejudice as something to be described, for example, in terms of how much prejudice exists in a certain group of individuals, or what proportion of people in a sample are highly prejudiced and what proportion are largely lacking in prejudice. Rather, they are likely to want to explain it, which means examining its causes. The researcher may seek to explain racial prejudice in terms of personal characteristics (such as levels of authoritarianism) or in terms of social characteristics (such as education, or social mobility experiences). In reports of research you will often come across the idea of 'independent' and 'dependent' variables, which reflect the tendency to think in terms of causes and effects. Racial prejudice might be regarded as the dependent variable, which is to be explained, and authoritarianism as an independent variable, and which therefore has a causal influence upon prejudice.

When an experimental design is being employed, the independent variable is the variable that is manipulated. There is little ambiguity about the direction of causal influence. However, with cross-sectional designs of the kind used in most social survey research, there is ambiguity about the direction of causal influence in that data concerning variables are simultaneously collected. Therefore, we cannot say that an independent variable precedes the dependent one. To refer to independent and dependent variables in the context of cross-sectional designs, we must *infer* that one causes the other, as in the example concerning authoritarianism and racial prejudice in the previous paragraph. We must draw on common sense or theoretical ideas to infer the likely temporal precedence of variables. However, there is always the risk that the inference will be wrong (see Research in focus 27.6, for an example of this possibility).

The concern about causality is reflected in the preoccupation with internal validity that was referred to in Chapter 3. There it was noted that a criterion of good quantitative research is frequently the extent to which there is confidence in the researcher's causal inferences. Research that exhibits the characteristics of an experimental design is often more highly valued than cross-sectional research, because of the greater confidence that can be enjoyed in the causal findings associated with the former. For their part, quantitative researchers who employ cross-sectional designs are invariably concerned to develop techniques that will allow causal inferences to be made. Moreover, the rise of longitudinal research like the BHPS almost certainly reflects a desire on the part of quantitative researchers to improve their ability to generate findings that permit a causal interpretation.

Generalization

In quantitative research the researcher is usually concerned to be able to say that his or her findings can be generalized beyond the confines of the particular context in which the research was conducted. Thus, if a study of racial prejudice is carried out by a questionnaire with a number of people who answer the questions, we often want to say that the results can apply to individuals other than those who responded in the study. This concern reveals itself in social survey research in the attention that is often given to the question of how one can create a representative sample. Given that it is rarely feasible to send questionnaires to or interview whole populations (such as all members of a town, or the whole population of a country, or all members of an organization), we have to sample. However, we will want the sample to be as representative as possible in order to be able to say that the results are not unique to the particular group upon whom the research was conducted; in other words, we want to be able to generalize the findings beyond the cases (for example, the people) that make up the sample. The preoccupation with generalization can be viewed as an attempt to develop the lawlike findings of the natural sciences.

Probability sampling, which will be explored in Chapter 8, is the main way in which researchers seek to generate a representative sample. This procedure largely eliminates bias from the selection of a sample by using a process of random selection. The use of a random selection process does not guarantee a representative sample, because, as will be seen in Chapter 8, there are factors that operate over and above the selection system used that can jeopardize the representativeness of a sample. A related consideration here is this: even if we did have a representative sample, what would it be representative *of*? The simple answer is that it will be representative of the population from which it was selected. This is certainly the answer that sampling theory gives us. Strictly speaking, we cannot generalize beyond that population. This means that, if the members of the population from which a sample is taken are all inhabitants of a town, city, or region, or are all members of an organization, we can generalize only to the inhabitants or members of the town, city, region, or organization. But it is very tempting to see the findings as having a more pervasive applicability, so that, even if the sample were selected from a large city like Birmingham, the findings would be relevant to all similar cities. We should not make inferences beyond the population from which the sample was selected, but researchers frequently do so. The concern to be able to generalize is often so deeply ingrained that the limits to the generalizability of findings are frequently forgotten or sidestepped.

Student experience
Generalizability in a student project

For his team-based survey research on students at his university, Joe Thompson felt that issues to do with reliability and validity were important. In particular, it appears from the following comment that the generalizability of the findings was especially significant.

> Again, the main considerations were reliability and validity of the research. Thus the methods used reflected this; the questionnaire went through a modification period where we as a group not only tested it on our sample but also received information from staff who worked within the area our research project was aimed at. We knew that the sample had to be representative of the whole university, so the number of members from the group interviewing students from different halls was in ratio to the number of students who lived within those residences.

To read more about Joe's research experiences, go to the Online Resource Centre that accompanies this book at: ***www.oxfordtextbooks.co.uk/orc/brymansrm4e/***

The concern with generalizability or external validity is particularly strong among quantitative researchers using cross-sectional and longitudinal designs. There is a concern about generalizability among experimental research, as the discussion of external validity in Chapter 3 suggested, but users of this research design usually give greater attention to internal validity issues.

Replication

The natural sciences are often depicted as wishing to reduce to a bare minimum the contaminating influence of the scientist's biases and values. The results of a piece of research should be unaffected by the researcher's special characteristics or expectations or whatever. If biases and lack of objectivity were pervasive, the claims of the natural sciences to provide a definitive picture of the world would be seriously undermined. As a check upon the influence of these potentially damaging problems, scientists may seek to replicate—that is, to reproduce—each other's experiments. If there was a failure to replicate, so that a scientist's findings repeatedly could not be reproduced, serious questions would be raised about the validity of his or her findings. Consequently, scientists often attempt to be highly explicit about their procedures so that an experiment is capable of replication. Likewise, quantitative researchers in the social sciences often regard replication, or more precisely the ability to replicate, as an important ingredient of their activity. It is easy to see why: the possibility of a lack of objectivity and of the intrusion of the researcher's values would appear to be much greater when examining the social world than when the natural scientist investigates the natural order. Consequently, it is often regarded as important that the researcher spells out clearly his or her procedures so that they can be replicated by others, even if the research does not end up being replicated.

Whether research is in practice replicated is another matter. Replication is not a high-status activity in the natural and social sciences, because it is often regarded as a pedestrian and uninspiring pursuit. It is striking that, in the example referred to in Research in focus 7.7, the exercise is referred to as a 'replication *and extension* of several previous studies' (emphasis added), conveying the impression that it is not *just* a replication.

Moreover, standard replications do not form the basis for attractive articles, so far as many academic journal editors are concerned. Consequently, replications of research appear in print far less frequently than might be supposed. A further reason for the low incidence of published replications is that it is difficult to ensure in social science research that the conditions in a replication are precisely the same as those that pertained in an original study. So long as there is some ambiguity about the degree to which the conditions relating to a replication are the same as those in the initial study, any differences in findings may be attributable to the design of the replication rather than to some deficiency in the original study. To some extent, this is the case with the research referred to in Research in focus 7.7. Nonetheless, it is often regarded as crucial that the methods taken in generating a set of findings are made explicit, so that it is *possible* to replicate a piece of research. Thus, it is replic*ability* that is often regarded as an important quality of quantitative research.

Research in focus 7.7
Replicating a study of cartoons

S. N. Davis (2003: 412) conducted what she refers to as a 'replication and extension of several previous studies'. The replication was of previous research—particularly that of L. Smith (1994)—that conducted content analyses of the characters in commercial cartoons that are broadcast in between children's television programmes in the USA. Content analysis is a technique that aims to provide quantitative analyses of different kinds of content in a systematic fashion. It is covered in detail in Chapter 13. Davis (2003) was especially interested in the extent to which the cartoon characters exhibited sex-role stereotyping. Based on previous research, Davis deduced several hypotheses concerning the sex-role stereotyping of the cartoon characters in the 1990s. Examples of such hypotheses are:

- 'characters in major roles will be more likely to be male than characters in minor roles' (2003: 411);
- 'the character will be more likely to be male if the activity is an individual activity than a group activity' (2003: 411);
- 'characters in activities with high amounts of movement will be more likely to be male than those characters who are portrayed with low amounts of movement' (2003: 411).

Davis depicts her research as partly a replication and partly an extension, because Smith's research was concerned with children's television progammes in general, whereas hers is just concerned with animated cartoon programmes. She analysed the content of cartoons shown in one month of 1995. A cartoon entered the sample just once, no matter how many times it was shown. Through this process, there were 167 cartoons and 478 characters that were analysed. Her findings confirmed that advertising through cartoons aimed at children does indeed entail sex-role stereotyping. However, she also writes:

> As this project largely replicated Smith's research, it shows the need for continued replication of this kind of analysis, as some of the findings in Smith's research were not reproduced in the analysis. The differences could be a function of the more narrowly defined sample of television programming from which the advertisements were drawn, or they could show a change in advertisers' methods of advertising their products to children. (S. N. Davis 2003: 421)

This research shows that a replication can be very valuable in establishing that the findings from a study should not be too readily accepted at face value. On the other hand, the second sentence in this quotation demonstrates how difficult it is to interpret the findings of a replication study. It is difficult to know how to interpret any divergences in the findings. Instead, as Davis implies, it is not that the original findings are 'wrong' but that it could be that, when applied to a different kind of sample, the same kind of analysis yields different findings or that there has been a change in advertisers' practices. It is common for there to be ambiguities of this kind with replications in social research.

The critique of quantitative research

Over the years, quantitative research along with its epistemological and ontological foundations has been the focus of a great deal of criticism, particularly from exponents and spokespersons of qualitative research. To a very large extent, it is difficult to distinguish between different kinds of criticism when reflecting on the different critical points that have been proffered. These include: criticisms of quantitative research in general as a research strategy; criticisms of the epistemological and ontological foundations of quantitative research; and criticisms of specific methods and research designs with which quantitative research is associated.

Criticisms of quantitative research

To give a flavour of the critique of quantitative research, four criticisms will be covered briefly.

1. *Quantitative researchers fail to distinguish people and social institutions from 'the world of nature'.* The phrase 'the world of nature' is from the writings of Schutz (1962) and the specific quotation from which it has been taken can be found on page 13 above. Schutz and other phenomenologists charge social scientists who employ a natural science model with treating the social world as if it were no different from the natural order. In so doing, they draw attention to one of positivism's central tenets—namely, that the principles of the scientific method can and should be applied to all phenomena that are the focus of investigation. As Schutz argues, this tactic is essentially to imply that this means turning a blind eye to the differences between the social and the natural world. More particularly, as was observed in Chapter 2, it therefore means ignoring and riding roughshod over the fact that people interpret the world around them, whereas this capacity for self-reflection cannot be found among the objects of the natural sciences ('molecules, atoms, and electrons', as Schutz put it).
2. *The measurement process possesses an artificial and spurious sense of precision and accuracy.* There are a number of aspects to this criticism. For one thing, it has been argued that the connection between the measures developed by social scientists and the concepts they are supposed to be revealing is assumed rather than real; hence, Cicourel's (1964) notion of 'measurement by fiat'. Testing for validity in the manner described in the previous section cannot really address this problem, because the very tests

themselves entail measurement by fiat. A further way in which the measurement process is regarded by writers like Cicourel as flawed is that it presumes that when, for example, members of a sample respond to a question on a questionnaire (which is itself taken to be an indicator of a concept), they interpret the key terms in the question similarly. For many writers, respondents simply do not interpret such terms similarly. An often used reaction to this problem is to use questions with fixed-choice answers, but this approach merely provides 'a solution to the problem of meaning by simply ignoring it' (Cicourel 1964: 108).

3. *The reliance on instruments and procedures hinders the connection between research and everyday life.* This issue relates to the question of ecological validity that was raised in Chapter 3. Many methods of quantitative research rely heavily on administering research instruments to subjects (such as structured interviews and self-completion questionnaires) or on controlling situations to determine their effects (such as in experiments). However, as Cicourel (1982) asks, how do we know if survey respondents have the requisite knowledge to answer a question or whether they are similar in their sense of the topic being important to them in their everyday lives? Thus, if respondents answer a set of questions designed to measure racial prejudice, can we be sure that they are equally aware of what it is and what its manifestations are and can we be sure that it is of equal concern to them in the ways in which it connects with everyday life? One can go even further and ask how well their answers relate to their everyday lives. People may answer a question designed to measure racial prejudice, but respondents' actual behaviour may be at variance with their answers (Thinking deeply 12.2).

4. *The analysis of relationships between variables creates a static view of social life that is independent of people's lives.* Blumer (1956: 685) argued that studies that aim to bring out the relationships between variables omit 'the process of interpretation or definition that goes on in human groups'. This means that, for example, we do not know how an apparent relationship between two or more variables has been produced by the people on whom the research was conducted. This criticism incorporates the first and third criticisms that have been referred to—that the meaning of events to individuals is ignored and that we do not know how such findings connect to everyday contexts—but adds a further element—namely, that it creates a sense of a static social world that is separate from the individuals who make it up. In other words, quantitative research is seen as carrying an objectivist ontology that reifies the social world.

We can see in these criticisms the application of a set of concerns associated with a qualitative research strategy that reveals the combination of an interpretivist epistemological orientation (an emphasis on meaning from the individual's point of view) and a constructionist ontology (an emphasis on viewing the social world as the product of individuals rather than as something beyond them). The criticisms may appear very damning, but, as we will see in Chapter 17, quantitative researchers have a powerful battery of criticisms of qualitative research in their arsenal as well!

Is it always like this?

One of the problems with characterizing any research strategy, research design, or research method is that to a certain extent one is always outlining an ideal-typical approach. In other words, one tends to create something that represents that strategy, design, or method, but that may not be reflected in its entirety in research practice. This gap between the ideal type and actual practice can arise as a result of at least two major considerations. First, it arises because those of us who write about and teach research methods cannot cover every eventuality that can arise in the process of social research, so that we tend to provide accounts of the research process that draw upon common features. Thus, a model of the process of quantitative research, such as that provided in Figure 7.1, should be thought of as a general *tendency* rather than as a definitive description of all quantitative research. A second reason why the gap can arise is that, to a very large extent when writing about and teaching research methods, we are essentially providing an account of *good practice*. The fact of the matter is that these practices are often not followed in the published research that students are likely to encounter in the substantive

courses that they will be taking. This failure to follow the procedures associated with good practice is not necessarily due to incompetence on the part of social researchers (though in some cases it can be!), but is much more likely to be associated with matters of time, cost, and feasibility—in other words, the pragmatic concerns that cannot be avoided when one does social research.

Reverse operationism

As an example of the first source of the gap between the ideal type and actual research practice we can take the case of something that I have referred to as 'reverse operationism' (Bryman 1988a: 28). The model of the process of quantitative research in Figure 7.1 implies that concepts are specified and measures are then provided for them. As we have noted, this means that indicators must be devised. This is the basis of the idea of **operationism** or **operationalism**, a term that derives from physics (Bridgman 1927), and that implies a deductive view of how research should proceed. However, this view of research neglects the fact that measurement can entail much more of an inductive element than Figure 7.1 implies. Sometimes, measures are developed that in turn lead to conceptualization. One way in which this can occur is when a statistical technique known as *factor analysis* is employed (see Key concept 7.6). In order to measure the concept of 'charismatic leadership', a term that owes a great deal to Weber's (1947) notion of charismatic authority, Conger and Kanungo (1998) generated twenty-five items to provide a multiple-item measure of the concept. These items derived from their reading of existing theory and research on the subject, particularly in connection with charismatic leadership in organizations. When the items were administered to a sample of respondents and the results were factor analysed, it was found that the items bunched around six factors, each of which, to all intents and purposes, represents a dimension of the concept of charismatic leadership:

1. strategic vision and articulation behaviour;
2. sensitivity to the environment;
3. unconventional behaviour;
4. personal risk;
5. sensitivity to organizational members' needs;
6. action orientation away from the maintenance of the status quo.

The point to note is that these six dimensions were not specified at the outset: the link between conceptualization and measurement was an inductive one. Nor is this an unusual situation so far as research is concerned (Bryman 1988a: 26–8).

Key concept 7.6
What is factor analysis?

Factor analysis is employed in relation to multiple-indicator measures to determine whether groups of indicators tend to bunch together to form distinct clusters, referred to as factors. Its main goal is to reduce the number of variables with which the researcher needs to deal. It is used in relation to multiple-item measures, like Likert scales, to see how far there is an inherent structure to the large number of items that often make up such measures. Researchers sometimes use factor analysis to establish whether the dimensions of a measure that they expect to exist can be confirmed. The clusters of items that are revealed by a factor analysis need to be given names (for example, innovation and flexibility or autonomy in the example in Research in focus 7.6). It is a complex technique that is beyond the level at which this book is pitched (see Bryman and Cramer 2011: ch. 13), but it has considerable significance for the development of measures in many social scientific fields.

Reliability and validity testing

The second reason why the gap between the ideal type and actual research practice can arise is because researchers do not follow some of the recommended practices. A classic case of this tendency is that, while, as in the present chapter, much time and effort are expended on the articulation of the ways in which the reliability and validity of measures should be determined, often these procedures are not followed. There is evidence from analyses of published quantitative research in organization studies (Podsakoff and Dalton 1987), a field that

draws extensively on ideas and methods used in the social sciences, that writers rarely report tests of the stability of their measures and even more rarely report evidence of validity (only 3 per cent of articles provided information about measurement validity). A large proportion of articles used Cronbach's alpha, but, since this device is relevant only to multiple-item measures, because it gauges internal consistency, the stability and validity of many measures that are employed in the field of organization studies are unknown. This is not to say that the measures are necessarily *un*stable and *in*valid, but that we simply do not know. The reasons why the procedures for determining stability and validity are rarely used are almost certainly the cost and time that are likely to be involved. Researchers tend to be concerned with substantive issues and are less than enthusiastic about engaging in the kind of development work that would be required for a thoroughgoing determination of measurement quality. However, what this means is that Cicourel's (1964) previously cited remark about much measurement in sociology being 'measurement by fiat' has considerable weight.

The remarks on the lack of assessment of the quality of measurement should not be taken as a justification for readers to neglect this phase in their work. My aim is merely to draw attention to some of the ways in which practices described in this book are not always followed and to suggest some reasons why they are not followed.

Sampling

A similar point can be made in relation to sampling, which will be covered in the next chapter. As we will see, good practice is strongly associated with *random* or *probability sampling*. However, quite a lot of research is based on **non-probability samples**—that is, samples that have not been selected in terms of the principles of probability sampling, to be discussed in Chapter 8. Sometimes the use of non-probability samples will be due to the impossibility or extreme difficulty of obtaining **probability samples**. Yet another reason is that the time and cost involved in securing a probability sample are too great relative to the level of resources available. And yet a third reason is that sometimes the opportunity to study a certain group presents itself and represents too good an opportunity to miss. Again, such considerations should not be viewed as a justification and hence a set of reasons for ignoring the principles of sampling to be examined in the next chapter, not least because not following the principles of probability sampling carries implications for the kind of statistical analysis that can be employed (see Chapter 15). Instead, my purpose as before is to draw attention to the ways in which gaps between recommendations about good practice and actual research practice can arise.

Key points

- Quantitative research can be characterized as a linear series of steps moving from theory to conclusions, but the process described in Figure 7.1 is an ideal type from which there are many departures.
- The measurement process in quantitative research entails the search for indicators.
- Establishing the reliability and validity of measures is important for assessing their quality.
- Quantitative research can be characterized as exhibiting certain preoccupations, the most central of which are: measurement; causality; generalization; and replication.
- Quantitative research has been subjected to many criticisms by qualitative researchers. These criticisms tend to revolve around the view that a natural science model is inappropriate for studying the social world.

Questions for review

The main steps in quantitative research

- What are the main steps in quantitative research?
- To what extent do the main steps follow a strict sequence?
- Do the steps suggest a deductive or inductive approach to the relationship between theory and research?

Concepts and their measurement

- Why is measurement important for the quantitative researcher?
- What is the difference between a measure and an indicator?
- Why might multiple-indicator approaches to the measurement of concepts be preferable to those that rely on a single indicator?

Reliability and validity

- What are the main ways of thinking about the reliability of the measurement process? Is one form of reliability the most important?
- 'Whereas validity presupposes reliability, reliability does not presuppose validity.' Discuss.
- What are the main criteria for evaluating measurement validity?

The main preoccupations of quantitative researchers

- Outline the main preoccupations of quantitative researchers. What reasons can you give for their prominence?
- Why might replication be an important preoccupation among quantitative researchers, in spite of the tendency for replications in social research to be fairly rare?

The critique of quantitative research

- 'The crucial problem with quantitative research is the failure of its practitioners to address adequately the issue of meaning.' Discuss.
- How central is the adoption by quantitative researchers of a natural science model of conducting research to the critique by qualitative researchers of quantitative research?

Is it always like this?

- Why do social researchers sometimes not test the validity and/or reliability of measures that they employ?

Online Resource Centre

www.oxfordtextbooks.co.uk/orc/brymansrm4e/

Visit the Online Resource Centre that accompanies this book to enrich your understanding of the nature of quantitative research. Consult web links, test yourself using multiple choice questions, and gain further guidance and inspiration from the Student Researcher's Toolkit.