ONE

Inference and warrant in designing research

This chapter will:

- explain what is meant by methodology and how it differs from method;
- introduce the three main types of research question in social science, and how each is answered by drawing inferences from patterns found in data; and
- explain that methodology is always controversial, because all good things do not go together, and that trade-offs must be struck between the virtues of good research designs.

How does methodology differ from the study of methods?

Since this book is about methodology, we should start with that term. But, first, we should point out that many standard textbooks use it loosely to refer to anything to do with research methods. So do not be surprised if you read books or articles that claim to be about methodology, but which deal with issues that are excluded from this book. Our definition, by contrast, is narrow and specific, and distinguishes clearly between method and methodology in social science.

In this book, we define *method* as the set of techniques recognised by most social scientists as being appropriate for the creation, collection, coding, organisation and analysis of data.

- *Data creation methods* are used to produce the raw material of research, namely well-structured data – or sets of information – that can be used to perform further investigations, of the kind described below. Data creation methods include ethnographic or participant observation, focus groups, individual interviews, questionnaire surveys and so on.

- *Data collection methods* are procedures for capturing what is important for answering the research question from the data that have been created. They may involve scanning text for particular themes, codes or content or undertaking counts or more advanced quantitative procedures. However, we can only count or code once we have decided how to identify what is important, as we show in Example 1.1.

EXAMPLE 1.1. STREET LIFE

It is claimed that the number of people sleeping rough on the streets in British cities fell fairly sharply in the four or five years after the British Labour government's initiative on rough sleeping in 1998. But it then levelled out, and at the time of writing appears to be increasing again.

However, as a consultation paper issued by the subsequent coalition government (Department of Communities and Local Government, 2010) shows, there is a major problem with this claim, in that no one believes that the data on rough sleeping are accurate, because counting the number of rough sleepers is far from straightforward.

The government is worried that its official definition of rough sleepers as 'people sleeping or bedded down in the open air' means that local councils do not count people who spend the night awake or sitting up in sleeping bags. But does it follow that councils should count all people on the street with sleeping bags? What, for example, about people who may use them as an aid to begging, but do not actually sleep rough? And should councils count people who sleep in tents, stairwells of blocks of flats or who take refuge on cold nights in shelters run by charities?

- *Data coding methods* are procedures for determining whether the information indicated by a particular datum or set of data meet the standards or thresholds required for them to be classified under a category, where that category is related to the research question or hypothesis.
- *Data organisation methods* are procedures for laying out whole sets or series of data, that have either been created, collected and coded by the researcher for the purposes of the project, or been taken from another source – for example, a national survey data set such as the British Crime Survey (BCS) or, as in Example 1.1, the British government's annual estimate of rough sleepers (available at http://www.communities.gov.uk/publications/corporate/statistics/roughsleepingcount2010). Data organisation involves setting out the data on a suitably common basis – for example, by tabulating them – so that they can be analysed.
- *Data analysis methods* are procedures for manipulating data so that the research question can be answered, usually by identifying important patterns. Statistical procedures are obvious examples. There are many qualitative analysis techniques too, such as open-ended content analysis, and a variety of theory-based comparative techniques for handling historical qualitative data of the kind we shall discuss in Chapter 17.

EXERCISE 1.1. RIGOROUS BUT ROUGH

Taking account of the challenges identified in Example 1.1 above, think about how you could develop a consistent and accurate count of rough sleepers. What criteria would you use? How would you justify them?

Interpretation

You will notice that nothing has yet been said about the 'interpretation' of data. The process of data collection almost always requires the researcher to 'interpret' the data, and that this is particularly so when – as in Example 1.1 – the things being studied do not fall nearly into convenient, unambiguous units. We shall consider some of these issues in more detail when we discuss the use and application of concepts in Chapter 9. Coding likewise involves interpretation, because the decision whether the data indicate that a case meets standards for a particular code is an interpretive act of scientific judgement.

'Interpretation' is also required in the process of determining whether the data analysis supports the general conclusions drawn from the research, to answer the research question. We call this support, its *warrant*. Warrant is a central issue in methodology, and therefore one that will be addressed throughout this book.

A third meaning of 'interpretation' in methodology is discussed briefly below and will be discussed again in detail in Chapters 15 and 16. This third meaning of interpretation is restricted to particular kinds of data and particular sorts of conclusions – namely, those which attribute beliefs, ideas, emotions, or ways of classifying to people being studied.

But the point to emphasise here is that all methodological approaches rely to a large extent on 'data interpretation' and therefore 'interpretation' is not a separate stage or activity from the ones we list above. Although research proposals are often written with timetables describing 'data interpretation' as if it were the final stage of a project when conclusions are to be drawn about the theoretical or practical significance of the research, in fact interpretation is at the heart of the whole research process.

So what is 'methodology'?

The key lesson from this discussion is that *methodology* is not just – and is often *not very much at all* – a matter of method, in the sense of using appropriate techniques in the correct way. It is much more to do with how well we *argue* from the analyses of our data to draw and defend our conclusions. The *methodological* question posed by

our rough sleepers example is just what would allow us to claim that an increase in rough sleeping has occurred; that is, to make inference to a *description*. If we went on to claim that a rise in rough sleeping is being caused by the economic downturn, then this would be an inference to an *explanation*. Or perhaps it would be illuminating to explore what rough sleepers themselves would count as rough sleeping and why. This would require an inference to an *interpretation*.

Because methodology is about arguments that show warrant for inferences, it makes no sense to break down the study of methodology according to the different stages involved in the research process, in the way that we have just done above for methods. Rather, we shall distinguish in this book between different approaches to methodology, and discuss the strategies appropriate to these approaches. We shall begin this discussion later in this chapter, when we discuss the differences between research designed to lead, respectively, to *description*, *explanation* and *interpretation*. But we stress throughout the book that each of these approaches raises the same basic methodological question – how and how far can you argue from the particular data to the particular conclusions, or, to put it another way, what argument, if any, do these data actually support?

Being able to draw sound conclusion depends on designing all stages in the project on sound methodological principles. Conversely, it is entirely possible to follow prescribed methods carefully, but still produce methodologically suspect research, if the conclusions drawn from it are not soundly based. These problems are inescapably theoretical ones, because the study of methodology involves theories about how and how far the research design enables us to draw sound inferences to conclusions that provide answers to our research questions, or that determine how far our hypotheses are supported or undermined.

And that is what this book is all about.

Inference and warrant

The core concepts in methodology are those of inference and warrant, and we should explain here why they are so important.

We are used to opinion pollsters drawing conclusions about the voting preferences of over forty million electors by sampling the opinions of around a thousand people. They do this by using widely accepted principles of statistical inference. This example illustrates the problem that we often need to draw conclusions about a large population from what we can find out about a smaller sample. A second problem is that we cannot always observe the things we are interested in directly, but are forced to work with proxies or indicators. For example, psychologists make inferences about the working of human or animal brains from observing very fine movements of eyes. Industrial sociologists make inferences about organisational morale from the way workers behave or describe their feelings. And anthropologists interpret how human beings make sense of their worlds from their stories or other cultural artefacts. In none of these examples can synapses firing in brains, 'morale' or 'sense-making' be directly observed.

Furthermore, researchers could not confidently make inferences without theories – however implicit or provisional – about the relationships between the things in which they are interested and those things which they can directly observe. For example, using cultural artefacts to interpret sense-making depends on a theory of culture.

We can therefore define *inference* as (1) the process of making claims about one set of phenomena that cannot be directly observed (2) on the basis of what we know about a set of things that we have observed where (3) the choice of research instruments depends on a theory of how those instruments work.**[C1Q1]**

We can define *warrant* as the degree of confidence that we have in an inference's capability to deliver truths about the things we cannot observe directly. Warrant involves particular standards, which we shall discuss in more detail in subsequent chapters. We shall see, too, that some of these standards are more straightforwardly related to methods than others.

Observation

In the course of the book, we shall have occasion to use this slippery but absolutely unavoidable word in several ways. There are four different ways in which this word is used in social science methodology:

1 The value taken by a unit of data that is collected for, defined by and organised in a scheme of measurement. For example, the value ascribed to a variable entered into a cell on a spreadsheet or table is an observation on that variable. 'Observation' is used in this sense in the question, 'what do the observations show?'
2 A unit of data, such as a case in a sample or data set, as in the question, 'how many observations do you have?'
3 The systematic collection of data about behaviour or action, where the researcher cannot exercise experimental control over the regime of stimulus and constraint under which the research participants act, as in the term, 'observational research', which is the alternative to experimental research.
4 The activity of a researcher undertaking visual and/or audio inspection of participants' behaviour, as in 'a period of fieldwork observations'.

When we discuss some philosophical questions in Chapters 2, 3 and 4, we shall use the word in sense 1 a good deal. Chapter 5 considers observational research, in sense 3. In Chapter 6, when we discuss variable-oriented research, 'observation' will be used in sense 2. Although we bear sense 4 in mind throughout, it will come to the fore particularly in Chapters 15 and 16. You are warned to pause whenever you see the word, to make sure that you know what is meant. It will always be clear from the context which meaning is intended, but you can check either this page or the entry in the glossary if you need a reminder.

Some controversial claims about methodology

With those definitions in mind, it is time for us to make some big claims. Some are going to be controversial. You will find, as you read this book, that almost anything that is said in the field of methodology will attract disagreement. This is another big difference from the study of methods, because most people who study methods agree on what counts as, for example, transcribing an interview, or calculating a chi-squared test.

Here is our first big claim. *Making warranted inferences is the whole point and the only point of doing social research*, irrespective of what type of data and what style of research we use. The contribution to knowledge of any research consists in the inferences that can be made from it. Inferences are the principal products; they provide support for findings; and they are what make findings into findings rather than speculations, on the one hand, or raw data, on the other.

There are two reasons for making this claim. The first is a semantic one and the second rests upon a normative claim about what our ambitions ought to be and why social scientists get out of bed in the morning.

The semantic reason is that careful attention to inference, and what warrants it, is what distinguishes *research* from other kinds of investigation. Good journalistic reportage does not generally try to make inferences, beyond telling us what the reporter found. Interesting speculative or theoretical writing does not have to be so concerned with warrant: pure theorists and social commentators leave that to empirical researchers. Detective work by police officers, however, *is* concerned with inference and warrant. But it differs from most social research – although it does resemble some kinds of historical work – in that it is concerned only with warranted inference about the particular case under examination, whereas a good deal of social science research is interested in drawing inferences beyond the particular case to a wider population.

The bigger, normative reason is that warranted inference is worth doing, because it represents a strategy for making a contribution to knowledge that none of these other investigative activities can achieve, given their entirely proper purposes and limitations. We need to understand how social processes generally work, and this cannot be done adequately by nailing 'whodunnit' in a particular criminal case, or by news-hounds ferreting out facts, or even by reading the insights of literary giants.

Critics of inferential ambition

We acknowledged that our first big claim would be controversial, and we should therefore tell you who would object to it, and why. Those who would resist this claim tend typically to argue one of the following positions:

- Social research can be justified if it 'gives voice' to people – such as rough sleepers – whose perspectives on homelessness would not otherwise be available. For this purpose, it is claimed, warranted inference to general theories is not necessary and, indeed, may

actually be harmful. What, rather, is needed is researcherly observation and analysis that is faithful to the views of the individuals studied.

- Social research cannot, because of its inherently subjective nature, achieve warrant for general inferences, and should be considered just as lacking in fundamental – or 'foundational' – warrant as journalism, speculative writing, *belles lettres* and detective work. On this view, the accounts, say, of rough sleepers, national government ministers, local charity managers and local mayors of why rough sleeping is a problem, and how big and significant a problem it is, are bound to be different: we cannot achieve a perfectly accurate description, let alone a true explanation, of this state of affairs. This view is shared by several schools of social thought, ranging from scepticism through relativism to anti-foundationalism and postmodernism.

An answer to the critics

We are not persuaded by either of these claims, and, for the record, we shall offer a couple of remarks to indicate why we disagree with the first of these views. The second, we shall leave for Chapters 2–4.

'Giving voice' involves attributing thoughts, emotions, practices, aspirations, memories and so on to other human beings. Researchers often want to reveal the preferences, experiences or ways of understanding the lives of the people they study. But none of these things can be observed directly, nor can they be read off unproblematically from what people say in interviews or do when they are observed. There is no getting away from using information from outside the particular situation, because even the concepts we use are taken from a wider vocabulary. And when we try to work out just what people think, we draw on information about other people we believe to be similar to those we are studying.

The very concept of 'rough sleeping', for example, is one drawn from government policy documents, and we would find it difficult to escape from concepts such as the 'vulnerability' or 'social exclusion' of rough sleepers if we tried to describe the impact on their lives of, say, the closure of a winter shelter due to spending cuts. People in charge of public policy, and researchers who write and read scholarly articles or monographs, use language in a very different way from many of those studied by social science research. And so, 'giving voice' often involves risky acts of translation or making risky attributions.

The only way to do it well – and to do it in ways that make us accountable to other academics or to participants in our research – is to adopt procedures that force us to be conscious of the inferences we make and to reveal all our workings-out. That is to say, one characteristic of good research design is that it enables us to demonstrate how we got from interviews and other observations to our conclusions about research participants' lives. This process *is* warranted inference.

An alternative way of proceeding is available, of course. We could write down what we happen to think and perhaps publish it in journalistic outlets. But our research

would then constitute a different kind of enquiry, undertaken for different purposes and with different kinds of accountabilities to the data, to research participants and to the wider academic community. Social research based on warranted inference makes a quite distinct – and distinctly valuable – contribution to understanding people from that made by journalism or any other type of enquiry. Specifically, the unique contribution of social science consists in the methodological care that we pay to the inferences we make.

Inference to what?

All this raises an important but obvious question: *to what* exactly do we make inferences? We have already seen that social scientists distinguish between three types of purposes for which inferences are made. These purposes are *description*, *explanation* and *interpretation*.

Descriptive inference

Descriptive inference is undertaken to answer certain questions about Xs (where X stands for any empirical topic for social research) when we cannot observe them at all, or cannot observe them all, or can observe only aspects of them, or cannot be sure that what we are observing of the Xs is quite what it seems. These questions are, 'what kinds of things are the Xs?', 'what kinds of statements can we make about them?' and 'how can we characterise them?' The product of descriptive inference is a set of claims about Xs. These claims may be about what is typical of Xs, what is generally true about Xs, or what is true about a subtype or across some spectrum of Xs.

One product of research on rough sleepers, for example, might be a description of how many rough sleepers there are in a particular town; what kind of people they are by age, gender and so on; how long, on average, they have been homeless; whether this period is becoming longer or shorter; and whether the number of long-term homeless people is rising or falling. This description would depend on inference, because – even if we could count directly everyone who sleeps rough in the town on a particular night – we would need to make assumptions about what proportion of rough sleepers we have observed. And we shall also have to make inferences from earlier data or from interviews with rough sleepers, about *changes* in patterns of rough sleeping and in the characteristics of the population of rough sleepers.

Some textbooks are very snooty about descriptive inference. This snootiness is – to coin a phrase – unwarranted. Description may be a modest ambition, but it is a necessary one. It is very difficult to go on to do anything more ambitious in social research if you have not got the descriptive inferences right. It is true that the most prestigious journals do not publish articles that offer *only* descriptive inferences. But the articles that they do publish rely, in a vital part of their overall argument, on the soundness of descriptive inferences, even if those parts of their workings are not shown.

Explanatory and counterfactual inference

Explanatory inferences are undertaken to answer the questions, 'why have the Xs done Z or become Y?', 'what brought this about?' and 'what *caused* the Xs to become Y or do Z?' In Chapters 10–13, we shall look in much more detail at what we understand by causation. We shall see that explaining how something came about raises methodological challenges of a higher order than describing it, although description can often be quite tricky too.

Suppose that we want to find out whether cuts in public spending have contributed, causally, to an increase in rough sleeping. Once the cuts have taken place, we can no longer look at rough sleeping in a particular town in the absence of those cuts. So we could never measure the impact of cuts on rough sleeping by comparing the situation we currently observe with one (in one and the same place) in which the cuts had never taken place. This difficulty is known as the fundamental problem of counterfactual causal inference. It is one reason why explanatory inference is tough. But it is often very important to try for explanations. Indeed, explaining why things happen is the main reason that anyone pays for social science research to be done, in the hope that explanation will help with the design of interventions in social problems.

There are weaker senses of the term 'explanation' which do not require *causes* to be revealed. For example, researchers write of *statistical* 'explanation'. This phrase refers to the process of showing that two variables are strongly associated with each other, but does not require us to draw any inferences about which direction any influence might run or to rule out the possibility that both variables are being influenced by a third variable. Other explanations are *logical* in character. That is, we may 'explain' a condition or event, by showing how it is derived logically from another. For example, we may explain the government's plans for extended sharing of personal information about individual citizens between government agencies on the grounds that this is a direct – that is, a logical – implication of an emphasis on multi-agency interventions in social problems such as homelessness.

Interpretive inference

Finally, there are *interpretive inferences*. Interpretive inference is addressed to a variety of questions, some of which we have already discussed.

We have seen that the most elementary interpretive inference is made when we determine whether something is to count, for a given research purpose, as falling within some category, and therefore decide that it is to be given a particular code or measure. We call this an interpretation of its *categorical significance*. Deciding, for example, who counts as a rough sleeper – a question which precedes the descriptive inference question of whether we can draw conclusions about the numbers of rough sleeping – is clearly a matter of interpreting the concept of a rough sleeper. And that, in turn, depends on our view of whether that concept captures the particular aspects of the underlying condition of utter homelessness in which we are interested.

Second, giving voice is only one way of accounting for how people think, feel, understand, frame issues and so on. Interpretive inference is not simply the development of descriptions of people's subjective experiences, but may also produce an *integrated* account – or interpretation – of the *subjective significance* for people's mental lives, in which the patterns observed make some larger sense. For example, in interpreting how managers of local council housing departments perceive the implications of the government's edict to count rough sleepers in new ways, we would probably need to go beyond a simple repetition of our descriptive data (e.g. 37% of respondents agreed with Proposal 1 put to them in our survey) by drawing inferences about the significance they attach to the government's proposals for the lives of their rough sleepers and for their capacity to help them. If the data allow, we could also, perhaps, make further inferences about what these managers believe to be the significance of these proposals more broadly for social justice or social inclusion and about the standards they appear implicitly to adopt in measuring justice and inclusion. Finally, there are inferences to integrated accounts in which the subject of the interpretation is not the mental life of a group – or groups – of people, but a set of events. Historical interpretation – which is very important in historical sociology, comparative political science, business history and even in institutional economics – is a case in point. Its aim is to detect overarching patterns of historical events – for example, those involved in the emergence of multinational business corporations in the period after the First World War or in the growth of a welfare state after 1945 – to provide the basis for an integrated account or interpretation of their *objective significance*.

Relationships between descriptive, explanatory and interpretive inferences

Much of the book will be devoted to considering separately the standards of warrant required for inferences in *explanatory* and in *interpretive* research. But we shall see that almost all explanatory and interpretive research rests upon descriptive and categorical interpretive inferences.

We shall see, too, that even those researchers who insist most fiercely on an exclusive focus on interpretation of *subjective* significance cannot, in practice, carry out that task without implying some kind of explanation of why people think as they do. It is very difficult to develop an account of, say, the ways in which people on low incomes think about or 'frame' the risks of health problems arising from their diets, without making some reference to categories that imply causation. For example, in trying to decide between interpretations that emphasise the limited dietary choices available to people on low incomes and those interpretations which emphasise their limited willingness seriously to consider eating health foods, researchers necessarily find themselves implying something about the causal role that beliefs might play in explaining unhealthy dietary behaviour.

This example illustrates the point that separating descriptive from causally explanatory categories is not straightforward, because we often describe by using categories that imply an explanation. For example, we might count the number of 'drug-dependent' people who are registered for treatment in the UK, but the very use of this category recognises addictive dependency as a significant *cause* of the use of illegal drugs and carries the implied claim that it should be treated rather than punished. We shall see in Chapter 15 and 16 that there are other and deeper reasons why it is difficult to do interpretative research without carrying any explanatory baggage.

The questions addressed by descriptive, explanatory and interpretive research are, nevertheless, analytically quite distinct. These three types of research ask, respectively, 'what's going on with the Xs?', 'why have the Xs done Y?', 'what do the Xs understand by the way they do Y?' and 'what is the wider significance of the fact that the Xs have done Y?'. It is therefore most helpful to consider separately the methodological challenges raised by each of these three approaches, and this is what we shall do in this book.

Trade-offs between virtues in warranting inference

In examining these challenges, we shall explore the virtues that should be exhibited by methodologically sound research, if it is to warrant the inferences that it seeks to support. Indeed, we have already noticed some of these virtues.

First, in discussing description, we have implied that a key virtue of a description is that it should be as *accurate* as possible within the limitations imposed by the ways in which the data have been created and collected. For example, the kind of accuracy we expect from a statistical description of broad trends is very different from the kind that can be achieved by a meticulous anthropologist who carefully checks each significant observation recorded in his or her field notes.

Second, we have assumed that our inferences should capture the significance of as many of the data in the set as is practicable. In other words, the account should summarise and integrate our findings, but with the minimum loss of the facts, nuances, differences and contrasts that are relevant to the question. The better our account does this, the better its *goodness of fit*.

Third, in contrasting social research with detective work and investigative journalism, we pointed out that social science researchers want to draw inferences beyond the particular case to some wider population of people, events or cases. That is, we are interested in achieving *generality* across some category.

Fourth, we mentioned that researchers often look for a few overarching patterns that are of the greatest significance in shaping thought styles or emotions or in explaining outcomes or events. Whilst it might be tempting to trace in great detail the interaction of a large number of complex factors that might *explain*, say, the rise in custodial sentences handed down by the criminal courts, it is both impractical and distracting to continue piling up lots of different factors over a large number of cases.

It may be better to compare the influence of a few, important factors such as changes in national sentencing guidelines, judges' attitudes in interpreting them, and beliefs held by judges and juries about how community sentences work. That is, another virtue of both explanations and of interpretations is *parsimony*.

There are other virtues, which we shall consider in due course. However, we shall also see that it is often impossible in the same research design to maximise accuracy, goodness of fit, generality and parsimony, let alone other virtues (Przeworski and Teune, 1970). For example, the more accurate we try to be, the more detail we accumulate and the closer we stick to the granularity of particular cases, the more difficult it becomes to generalise across cases. It also becomes more difficult to identify the effects of a few really central factors, because they will not consistently perform their explanatory or their interpretive work at the level of close detail. Conversely, the more parsimonious we want to be, the more likely it is that we shall be forced to restrict the domain of cases over which we can generalise, because rather few things are common to every case, especially those falling in wide categories like 'homelessness' or 'judicial behaviour'. This problem means that we have to strike trade-offs between virtues in designing our research.

The need to strike trade-offs between virtues of good research design is one reason why there is no such thing as a piece of research that is completely beyond methodological impugning. It is possible to complain about something in every piece of social research, and social scientists, being a quarrelsome lot, are not slow to find it. But that does not mean that anything goes in striking trade-offs. There are always better and worse trade-offs to be found to address a particular research question, and there are some that lie so far behind the trade-off curve, or so far to one extreme of that curve, that they would clearly constitute poor research design.

What is 'research design'?

But what, actually, is 'research design'? By the *design* of a research project, social scientists usually mean (1) the specification of the way in which data will be created, collected, constructed, coded, analysed and interpreted (2) to enable the researcher to draw warranted descriptive, explanatory or interpretive inferences (3) where the warrant is calculated to strike a reasonable trade-off between competing virtues; and (4) where the standards of warrant may vary slightly, but are based on a core set of virtues for each type of inference.

A research design is usually set out in advance of undertaking a project, in a research plan or *proposal*. A more detailed statement of the methodological defence of a research proposal is often provided in a *protocol*, which lays out in detail the steps through which the inference will proceed and the degree to which the conclusion can be supported, given the nature of the data and the nature of the methods used to create, collect, code, construct and analyse them.

Standards of good research design

The simplest standards of soundness in methodology are those of *reliability* and *validity*.

Reliability

Reliability has to do with how we measure – or, if you are using qualitative data, code – the things in which we are interested. A reliable system of measurement or coding is *consistent* in that, each time it is used on the same data, it yields the same measure or code. If two researchers work together, and both follow the same procedure on the same data, they should produce the same measures or codes. Redoing the coding or measurement, to see how reliable the procedure is, is called the 'test/retest' method of assessing reliability.

A second way to assess reliability at the level of method is called the 'internal consistency' method. This does not rely on repeating the coding or measurement of the same data, but on gathering additional data using the same design. In a questionnaire survey, for example, we might insert several questions, each phrased slightly differently, to ask the same thing. If they elicit the same answers from the respondents as did the first, then they provide some evidence that the first question was reliable.

Validity

Validity is, loosely, the degree to which our statements approximate to truth. It is conventional to distinguish between construct and conclusion validity, and between internal and external validity.

Construct validity is the degree to which the measures or codes used to operationalise a concept really capture what we intend to capture. For example, suppose we want to know how much 'goodwill' people have toward their neighbours in their own street. Goodwill is not a straightforward concept. We might ask about people's attitudes to other people in general and to their neighbours in particular. Perhaps we should ask about hypothetical future neighbours who might differ in important ways – for example, in their ethnic origins – from the present ones. But we should surely want to know, too, about how people actually behave toward different neighbours. Perhaps we would want to know about how they think they would behave in certain hypothetical situations, such as a severe fall of snow in the neighbourhood. We might also want to know about how they expect their neighbours to behave toward them.

Having settled on a set of measures or codes, we could assess their construct validity in several ways. The simplest way might be to look at theories of goodwill, and compare our measures or codes with the features used in those theories. If we are collecting quantitative data, we might use statistical analysis to determine whether there are common factors that run through each of our chosen measures – such as whether goodwill is based, say, on 'social affinity' (sharing ideals and beliefs) or

'social reciprocity' (helping each other out), or whether in our observed cases, the values point in different directions: for example, goodwill may be strong where it depends on affinity but less so where it depends on reciprocity. If so, we might wonder whether, in fact, 'goodwill' is a single phenomenon after all, and instead stipulate different 'types' of goodwill. This process would increase the construct validity of our concept of goodwill, by giving it more operational precision.

Measurement validity is a subtype of construct validity. It captures the extent to which any given measure or code allows us to attribute values, say to different factors in, or dimensions of, 'goodwill' without importing systematic bias. Measurement validity is important whether we use cardinal (1, 2, 3…) or ordinal (1st, 2nd, 3rd…) numbers, on–off codes (yes/no) or qualitative values (such as 'strong/moderate/weak').

Conclusion validity concerns the warrant we have for making inferences from our conclusions. It relates to the degree of support which the patterns observed in the data provide for the conclusions drawn from them. If we conclude, for example, that goodwill based in social affinity tends to be stronger than that based in social reciprocity, the question is whether this conclusion is a reasonable statement of what the data show.

Internal validity applies within a study, regardless of whether we want to generalise to others. It concerns the warrant we have for inferring that an outcome can be explained by a particular causal factor. If we claim, for example, that our study shows that 'social reciprocity' becomes stronger the longer people are neighbours, regardless of factors such as race, then the test of the study's internal validity is the extent to which we can show from our data that this really is the case.

External validity concerns the warrant we have for inferring that our findings would hold in other situations or studies that were similar, in relevant ways. Clearly, there is a gradient of similarity and dissimilarity. As samples or cases become less similar, external validity is bound to fall, along with our ability to generalise from the study. So, for example, the findings of a study of neighbourly goodwill in an American small town might be expected to hold in towns of a similar size and with a similar socio-demographic structure, but not in a city or in a small town with very different population. This means that a key issue in securing external validity is knowing what features of our cases or population are 'relevant' for this purpose, and what makes them 'similar' or dissimilar.

Trade-offs between validity and reliability

Just as there are trade-offs between different virtues in research design, so there can also be trade-offs between validity and reliability. At first blush, this might seem an odd claim. After all, as a measure or code declines in reliability, so it must also become less valid.

But there are some things we may want to measure or code in social science that are not amenable to straightforward measurement or coding. Suppose, for example, we want to understand the differences between people in respect of their capacity to make discriminating and thoughtful judgements in the fields of arts such as music, theatre, literature and dance. Measuring taste, or aesthetic judgement, requires a cluster of different dimensions, because it is not just one thing. We should need to measure or code the breadth of arts over which someone was capable of exercising judgement; whether they did so in consistent ways; and also the different ways in which they might be more and less articulate in their judgements; and so on.

Bringing all these measures or codes together into one composite indicator of taste could be done in a variety of different ways. We could, for example, increase the validity of our composite measure or code of taste by adding more subsidiary measures, such as scope, consistency and articulacy. That process would pick up more dimensions of this complex concept, but it would increase the difficulty of choosing a way of combining them, and our composite measure would be sensitive to whatever method we chose to weight and relate measures of particular dimensions of taste.

In other words, we would risk reliability for gains in validity. Beyond a certain point, too great a sacrifice of reliability will also ruin validity, and the range of *acceptable* trade-offs between the two values – for example, between reliable precision and valid relevance – especially in measuring complex or rich qualitative concepts such as taste is probably quite narrow. But there is usually more than one defensible trade-off to be struck dealing with this problem.

Sometimes, though, the trade-off problem can become vicious. Problems arise most acutely where the very process of doing the measurement or coding changes the thing being measured. For example, doing research about behaviours which are unlawful or which are regarded as immoral can cause the people being studied to behave more cautiously because they are being watched, or, alternatively, to show bravado by exaggerating their sinful behaviour. This is a problem that is well recognised, for example, among researchers who want to conduct ethnographic studies of institutional racism or bullying, where attempts directly to observe behaviour by means of non-participant observation end up by seriously undermining both validity and reliability.

This problem is also familiar to policy-makers. Goodhart's law was originally developed in the 1970s and 1980s, when new, more complex measures were developed by central banks of what counted as 'money'. The reason for measuring money in different ways was that central banks began to be charged with gaining control over the money supply, and needed to know how well they were doing. Unfortunately, the introduction of measurement and the use of policy measures to influence the money supply interacted in unexpected ways. Quite simply, when measures focused on one definition, people created money on some other definition instead: the central banks' work began to seem like squeezing a balloon in one part, only to expand the bulge at another.

The former Bank of England economist Charles Goodhart concluded that the very effort to measure money was making those measures less valid. He generalised his

finding to any situation in which measurement was associated with policy action and so had behavioural consequences. Goodhart's original formulation of his law concerned the application of policy action – 'any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes'. Later formulations have extended it to make the point that even introducing or publishing a measure will have behavioural consequences that reduce its validity for capturing the phenomenon of interest. The problem of Goodhart's law matters most in research conducted over a period of time, when the people being studied have time to react to the research. So it particularly affects longitudinal research or where the activity being studied is one about which people have normative views.

We shall return to these concepts in Chapter 6 to explore how they are applied to research designs that use variables. In the more advanced chapters about explanation and interpretation, we shall look at various ways in which internal validity can be pursued in observational research, not least because many method textbooks give only experimental examples of internal validity. Construct validity is of central concern in Chapter 9 on concept formation and is at the heart of the methodological challenges for good interpretive work that we discuss in Chapters 15 and 16.