01-Sapsford -3330-01.qxd 9/27/2005 3:22 PM Page 1

PART I
DESIGN ISSUES

1
Validating Evidence

Roger Sapsford and Victor Jupp

In this chapter we shall be looking at the issues which logically (and generally in practice) precede data collection itself – what cases to select, and how the study should be designed. The major concern is with *validity*, by which we mean the design of research to provide credible conclusions: whether the evidence which the research offers can bear the weight of the interpretation that is put on it.

Every report of research embodies an argument: 'on the basis of this evidence I argue that these conclusions are true'. Within this, the evidence presented is, again, the product of a series of arguments. The authors collect certain information in certain ways from or about certain people or settings, and the research report argues that this information may be interpreted in certain ways to lead to true conclusions about a certain population. What has to be established in order that the report's conclusions can be believed is that the arguments embodied in the report are *valid* ones: that the data *do* measure or characterize what the authors claim, and that the interpretations *do* follow from them. The structure of a piece of research determines the conclusions that can be drawn from it (and, more importantly, the conclusions that *should not* be drawn from it).

We shall argue, throughout this book, that the same questions have to be answered by research studies in widely different styles and arising from widely differing epistemological bases. Some research takes an essentially *positivistic* approach to the nature of knowledge about the social world: it takes the nature of the world as relatively unproblematic – the main problems being how to measure it adequately – and it emphasizes the neutrality and separateness of the researcher from that which is under investigation. Such work is typically *reductionist*: it seeks to explain the whole by measurement and correlation of the behaviour of parts or aspects of it. Typically, it is *quantitative*: it works by measurement and analysis of relationships between the resulting numbers; and, typically, it aspires to the methods of the natural sciences. Other studies may take a more *interactionist* perspective, looking at the meaning of situations and actions for people, conceived as something not fixed and determinate but negotiated from moment to moment. Studies from this perspective are more likely to be *naturalistic*, trying to eliminate the reactive effect of research procedures on what is studied, and to proceed by observation and participation in the situation as a whole or by relatively 'open' or 'unstructured' interviews with actors in it. Others again will proceed from a more *constructionist* perspective, regarding the space within which meanings are negotiated as a product of history and of social structure, rather than just of immediate negotiation. Work of this kind is likely to proceed by the analysis of interviews or written/printed texts for models of the social world, which are implicit in the text and give clues to the framework within which the writer or speaker is working. Other studies again may be more concerned with *reflexive* awareness and deal not with how things are but with how they might be – not with human nature, but with human capability, for example (see Stevens, 1995, for a discussion of this) – and these again are likely to proceed by analysis of what people say, and observation of what they do, in a holistic manner.

Whatever the form of the research and whatever its epistemological grounding, when reading a research report we shall be trying to assess whether the conclusions follow validly from the evidence. We shall therefore be asking who was researched, by what methods, and whether the logic of the comparisons made in the report is sufficient to bear the interpretation placed upon it.

Counting Cases: Measurement and Case Selection

A part of the argument in a research paper entails showing that the subjects or cases investigated can be taken as typical or representative of the population under investigation; the technical term for this question is *population validity*. A second obviously important topic is *validity of measurement*: the question of whether the measures which are used really do deliver what the researcher claims for them, or whether they give vague and error-ridden results, or even a competent measurement of something that turns out to be different from the researcher's claims. (In many ways the second of these is even more important than the first; there is no point in taking great care in selecting the sample if the measurements taken from it are uninterpretable!).

We shall begin our exploration of these two topics, in this section, with a consideration of government administrative statistics and three surveys around the topic area of crime and criminality. (The first also serves to remind us that not all research collects fresh data; valid research can equally be carried out on data already collected by someone else; see also Chapter 5). At the end of the section, however, we shall look at a study of a rather different kind to show that the same questions may validly be posed.

Examples from UK Government Statistics

Government statistics always look authoritative and are frequently presented as carrying authority; they are, after all, prepared by government statisticians and used by government ministers for planning purposes. In fact, however, they differ very much in their quality as evidence.

At one extreme, we might look at the statistics of births and deaths. Births *have* to be registered, both by the parent(s) and by the hospital or doctor; this includes not only live births but also stillbirths and abortions. Deaths *have* to be registered before a funeral can take place. There are presumably a very few births and deaths which remain concealed – cases of murder or infanticide – but we may reasonably take these statistics as virtually accurate counts of what is occurring, as an accurate representation of the 'population' of births and deaths.

At the other extreme, let us consider *Criminal Statistics*. These are published annually by the Home Office, and they contain (among other things) a supposedly complete count of crimes recorded by the police. We may be tempted to take this as a valid measure of what crimes have been committed during the year, but it would be a mistake to do so without further thought.

Activity 1.1 (allow 5 minutes at most)

Why might it be a mistake to take the count of crimes recorded by the police as a valid measure of crimes committed?

We know that some crimes are uncovered mostly by police action (e.g. road traffic offences), but that others (e.g. burglary, car theft, rape) depend mainly on private individuals reporting them to the police. There is, therefore, an element of personal discretion in what comes to the notice of the police in the first place. (We know from other evidence that most car thefts are reported, for example – for insurance purposes – but only a minority of rapes). Once the complaint reaches the police, the police officer may decide to record or not to record, depending on his or her assessment of whether a crime has actually been committed, or to initiate further investigation before recording. Thus, at least two levels of individual discretion are involved before an actual occurrence appears as a crime statistic; what appears in *Criminal Statistics* is not an unedited record, but one that has passed through a 'filtering process'. We would therefore handle 'crimes recorded by the police' with some caution as a measure of the 'population of crimes committed'; their validity in this respect is in doubt.

There are very many sets of official statistics that have this discretionary character – more than there are statistics that can safely be taken as neutral records of events. To take one more example, death statistics may be a valid and reliable measure, but the same cannot be said for the 'cause of death' tables. These depend on decisions made by GPs, hospital staff or coroners and are prone to human error and the uneven working of discretion as to what to record in cases of multiple causation.

Three Kinds of Crime Survey

Surveys of the victims of crime have been influential tools of criminological research, particularly in the 1980s and 1990s (for a good review of such surveys, see Walklate, 1989). Their popularity has been given impetus by official policies relating to law enforcement and also to communities taking responsibility for crime prevention. On a methodological front, further impetus was given by the widely held view of social researchers that official statistics on crime, such as those published in *Criminal Statistics*, failed to provide accurate measures of the true level of crime. Victim surveys involve the selection of a representative sample from the population. Questions are asked of sample members: whether they have ever been victims of crime within a specified period of time and whether they reported the event to the police. A major landmark in the development of victim surveys was that these large-scale surveys reported a much higher incidence of victimization than was recorded by the official statistics on crime.

Closer to home, Sparks and his colleagues carried out a much smaller and more localized survey in 1977, involving a representative sample of three areas of Inner London; this found that nearly half the sample claimed to have experienced actual or attempted crime in a 12-month period and reported an 11:1 ratio of victim-perceived to police-recorded crime. Subsequent surveys from Merseyside (Kinsey, 1986) and Islington (Jones et al., 1986; Crawford et al., 1990) have come up with broadly similar figures.

Activity 1.2 (10 minutes)

Two questions about validity of measurement:

1 What do you think are the strengths and weaknesses of victim surveys in measuring the extent of crime, compared with the official statistics?
2 If a victim survey asked whether the respondents had experienced a crime (a) in the last year and (b) in a 12-month period 10 years ago, would the comparison of these two sets of figures provide valid data for examining changes in victimization over time?

The faults of the official statistics are well known; as discussed above, *Criminal Statistics* records not what crimes were *committed* during the year, but what crimes were reported *and determined by a police officer to merit recording*. However, victim surveys are also not without their problems, which need to be taken into account when they are interpreted.

1 They are based on sample surveys, and often on relatively small samples, so there will inevitably be some degree of sampling error (see Chapter 2).
2 They involve asking people whether they have been the victims of crime, and the act of questioning sets up a social situation which may affect the answers: the

form of the question or the politics of the situation could lead respondents to lie (in either direction), or to stretch the truth to provide what they perceive as an answer acceptable to the researcher (see Chapter 4).

3 An element of judgement is involved: the survey has to take as given the respondent's judgement of what constitutes a crime. This may lead to some over-estimation of the 'true' amount of crime. For example, a broken window may be attributed to vandalism, whereas in fact it was caused by adverse weather conditions. In other cases, judgement may lead to an under-estimation. For example, in social situations and settings where violence is the norm, a victim of a beating may not view himself as the victim of a crime, even though a court of law might well judge otherwise if the case ever came to court. (The second is a more complex example than the first because judgement is required on both sides; the courts could as easily be misjudging the situation as the victim in terms of who started the fight and whether the degree of violence used was excessive).

4 The count depends on the respondent's memory. It is possible that some events will be forgotten. More likely, however, is that the count will be inflated by crimes which occurred outside the period of the question: if you ask for everything that happened during the year from March, it is not always easy for the respondent to remember whether a given incident occurred in April, March or even February of last year.

This problem of memory is obviously even more acute for questions about what happened 10 years ago. Our memories are reconstructed in the light of present concerns, and we tend to be vague about precisely when things happened where a long time period is involved. We would hesitate, therefore, to accept a contemporary 'one-shot' survey of this kind, which asks individuals to reflect on the past, as adequate evidence of changes over time.

Activity 1.3 (5 minutes)

See if you can think of two alternative ways of collecting data about crime victimization over time which would produce more valid data than the retrospective 'one-shot' survey.

One way around the problem of retrospective data collection is to examine findings from a succession of one-shot surveys carried out at different points of time: a *time-series design* or *trend design*. For a sound base of comparison, what we need are successive surveys which sample the same population, are constructed in the same way and are uniform in their definition and operationalization of variables (that is, the translation of the concept into something measurable). For example, the General Household Survey (carried out annually by GSO) has asked questions about burglary victimization in several years of its operation, allowing comparison over time for this limited category of crime. The most notable recent example, however, is provided by the British Crime

Survey (BCS; see Mayhew and Hough, 1982). This was carried out by the Home Office Research and Planning Unit in 1982, 1984 and 1988, and there are plans to continue it every four years. While the individuals questioned are not the same at each period of time, they are sampled in the same manner and asked the same or similar questions about the crimes of which they have been victims. The main analysis provides data on crimes of different kinds, reported and not reported to the police. As with other victim surveys, the BCS uncovers substantial under-reporting of crime; it also testifies to the great fear of crime among the general population, and particularly among women, older people and those living in inner-city areas – fear out of all proportion to the measurable likelihood of becoming a victim of crime. The analysis also provides insights into the respondents' reasons for not reporting crimes. Perceived triviality of the offence is by far the most important reason, but perceived lack of interest and impotence on the part of the police are also important. Given that the same questions are asked at each point of time, there is the facility for making comparisons over time and therefore the basis for making assertions about social changes and social trends.

However, if your interest is in change in *individuals* over time rather than the changes in *populations* over time, it is a bad strategy to take unrelated samples in two time periods; if you find changes, you cannot know whether people have changed or whether the population has recruited a new kind of member. Because different individuals are chosen at each sampling point, time-series designs are not appropriate for examining changes in individuals or for studying individual development. This is the hallmark of the *longitudinal* or *cohort* design. A celebrated British example is the National Child Development Study, which selected a national sample of children born in one week in 1947 and followed them until they were in their late twenties (Douglas, 1964, 1976). Within British criminology, and the study of delinquency and the causes of crime in particular, the longitudinal or cohort study is typified by the Cambridge Study in Delinquent Development, carried out by Donald West and his colleagues.

The aims and broad strategies of the Cambridge Study were very much influenced by the previous work of the Gluecks in the USA (Glueck and Glueck, 1950, 1962), which had indicated the important influence of early family socialization and family circumstances on who did or did not subsequently become delinquent, although, by the admission of its director, it 'began as a basic fact-finding venture without strong theoretical preconceptions and without much clear notion of where it might lead' (West, 1982: 4). However, it departed from its American predecessor by using a prospective (or longitudinal) rather than a retrospective ('one-shot' or cross-sectional) design. In other words, it involved examining which individuals, out of an initial sample, subsequently became convicted of delinquent acts, as opposed to studying retrospectively the backgrounds of those who had already been convicted. The reasons for this were that

> Research with established delinquents can be misleading. Once it is known that one is dealing with a delinquent, recollections and interpretations of his upbringing and previous behaviour may be biased towards a preconceived stereotype. Moreover, deviant attitudes may be the result rather than the cause of being convicted of an offence. (West, 1982: 3)

In 1961, a sample of 411 working-class boys aged about 8 was drawn from the registers of six state schools in a London area, reasonably close to the researchers'

London office and with a reasonably high delinquency rate. Girls were not included in the sample, and only 12 boys came from ethnic minority groups.

> In other words, it was an unremarkable and traditional white, British, urban, working-class sample. The findings are likely, therefore, to hold true of many similar places in southern England, but they may tell us nothing about delinquency in the middle classes or about delinquency among girls or among immigrant groups. (West, 1982: 8)

The sample members were contacted at six designated ages between 8 and 21, and sub-sections of the sample were purposively selected and re-interviewed at later ages (persistent recidivists, former recidivists not convicted of an offence for five years, and a random sample of non-delinquents). The main findings (but, for more detail, see West, 1969, 1982; West and Farrington, 1973, 1977) were that five clusters of factors were predictive of delinquency:

- coming from a low-income home
- coming from a large family
- having parents considered by social workers as having performed their child-raising unsatisfactorily
- having below-average intelligence
- having a parent with a criminal record

Activity 1.4 (allow 5 minutes)

What key features of longitudinal studies are exemplified in the Cambridge Study?

The Cambridge Study typifies longitudinal studies in a number of ways. First, it is *prospective* as opposed to retrospective, following 411 8-year-old boys through their teens and twenties. Secondly, in doing so it focuses on *individual development*, especially in relation to the generation of delinquent behaviour and subsequent criminal careers. Thirdly, the study is *descriptive* in that it describes individual development and change, but also *explanatory* in the way the analysis seeks to identify factors which can explain why some sample members became delinquents. Fourthly, the study seeks to be *predictive* by investigating how far delinquent and criminal behaviour can be predicted in advance. Finally, the study illustrates a feature of longitudinal research which has not been emphasized so far in this text: that it is often closely related to policy formation.

> The major policy implications of the Cambridge Study are that potential offenders can be identified at an early age and that offending might be prevented by training parents in effective child-rearing methods, preschool intellectual enrichment programmes, giving more economic resources to poor parents and providing juveniles with socially approved opportunities for excitement and risk-taking. (Farrington, 1989: 32)

During the summer of 1991 the then Home Secretary, Kenneth Baker, held a number of seminars with teachers, clergy, magistrates, social welfare professionals and others to discuss the rising crime rate and especially the problem of youth crime. Subsequently, he outlined, in general terms, policies to identify potential offenders at an early age and to introduce the kind of preventive measures described above. His decision was influenced not only by the seminar discussions but also by the conclusions of the Cambridge Study (see *Guardian*, 19 September 1991, p. 19).

Activity 1.5 (5 minutes)

Now list the strengths and weaknesses of longitudinal cohort studies and time-series designs, as you perceive them in the light of all you have read so far, especially with regard to collecting valid measurements.

Unlike one-shot cross-sectional designs, neither time-series designs nor longitudinal cohort studies are dependent on the collection of retrospective data in seeking to relate past experience to present-day attitudes and actions. Cohort studies go beyond this in being able to collect a wide range of data about a large number of variables at different stages of the same individual's life. As Douglas, a leading exponent of longitudinal surveys, points out:

> A cohort study allows the accumulation of a much larger number of variables, extending over a much wider area of knowledge, than would be possible in a cross-sectional study. This is of course because the collection can spread over many interviews. Moreover, information may be obtained at the most appropriate time: for example, information on job entry may be obtained when it occurs, even if this varies from one member of the sample to another. (Douglas, 1976: 18)

So longitudinal surveys allow the collection of more data, and also its collection at the most *appropriate* time. The enhanced validity of measurement consequent upon sampling through time adds to the *overall* validity – the overall plausibility of the argument. This is further enhanced by the fact that the explanatory value of cohort studies is greater: the longitudinal dimension provides direct evidence of a time-ordering of variables and so gives more credibility to causal inferences which link contemporary attitudes and actions to previous background, experiences and events.

On the negative side, however, longitudinal studies are very costly compared with cross-sectional studies or even time-series designs, and they produce results very slowly. They require key members of the research team to make a long-term commitment to the project, and they also require research funding bodies with patience and foresight as to the long-term benefits of such research. With regard to the sample, there is always a risk that members will change their attitudes or behaviours as a result of being part of the study. What is more, the sample runs the risk of being seriously depleted by drop-out over the years – known as 'sample attrition'. A survey of major North American longitudinal studies (Capaldi and Patterson, 1987) found that the average attrition rate was 17 per cent; one in six of the original respondents was lost to the survey on re-interview for a range of causes, including disinclination to be interviewed, death, emigration or simple failure to notify a new address. What matters here is not just the size of the drop-out but the question of whether the representativeness of the remaining sample is seriously affected. Research on children with severe adjustment problems (Cox et al., 1977) found that they tended to be over-represented among the drop-outs. West and Farrington (1973) found that parents who were uncooperative or reluctant to participate in the study were more likely to have boys who subsequently committed delinquent acts. In attitude studies, those who are most likely to have flexible attitudes are those most likely to be geographically mobile. In workplace studies which use volunteer informants, it is quite clear that those who volunteer readily very often have different things to say about the organization than those who are not much interested in participating in the study.

A further problem shared by all research with a time dimension to data collection is that variables collected at early stages may not anticipate theoretical developments at a later stage, with the result that crucial data may not have been collected. This is sometimes referred to as 'the problem of fading relevancy'.

> Because a long-term longitudinal study takes a long time to carry out, there is a risk that the theoretical framework that served as a basis for the design of the study, for the choice of variables and for the construction of indexes and so forth, has become obsolete by the time the data collection is complete and the results can be published. (Magnusson and Bergman, 1990: 25)

It is, of course, possible to introduce different variables at later stages of either a cohort or a time-series design, but there is no way that analysis of them can benefit from the strength of the designs, which is *the comparison over time*.

A final problem common to all quantitative surveys is precisely that they are quantitative. This sub-section of the chapter has tended to concentrate on sampling and the comparison of samples, but we also need to keep measurement and the validity of measurement in mind. In order to cover large numbers of cases, it is generally necessary to degrade information to numbers – 'readings' on scales – or to collect what data can readily be counted (i.e. *size* of family rather than what it is *like* to live in families of various sizes).

In the next sub-section we shall look at a very different kind of study which provides much richer and more 'naturalistic' data. We shall be suggesting that it has the same concerns as survey research, however, and equal though opposite problems. The focus will be on case selection – what we can say about the population on the basis of cases selected from it – and the extent to which valid conclusions can be drawn from what was said to the interviewers (a question akin to 'validity of measurement').

A Study of Mothers

In the early 1980s, Abbott and Sapsford (1987b) carried out an interview project in a 'new town', interviewing mothers of children with learning difficulties (at that time labelled as 'mentally handicapped'), plus a sample of mothers whose children were *not* identified as 'handicapped'.

> Sixteen families were contacted from a list extracted for us from the school rolls of two Special Schools in the new city (one designated for the mildly handicapped and one for the severely). We carried out two interviews with each mother, separated by about a year, not

> using a formal questionnaire but rather trying for the atmosphere of a friendly chat about life and work between neighbours. Although the interviews were tape-recorded, it seemed to us that this atmosphere was readily attained in most cases – the more so because Abbott was very evidently pregnant during the early interviews. … Although we cannot claim that [this study] had a sample statistically representative of the population of mentally handicapped children, we would claim that [it covers] a good part of the range – from the mildest of borderline handicaps to the very severe. These data are contrasted with a parallel series of interviews with mothers of children who have *not* been labelled as mentally handicapped. (Abbott and Sapsford, 1987b: 46)

The target sample was obtained from two schools for children with learning difficulties, which between them covered the whole range. In one, the researchers picked names at random from the role and wrote to them, building up a sample from those who replied and were prepared to be interviewed. In the other, the head teacher wrote to families (excluding those she knew to be in substantial distress at that time and one or two who were known to be uncooperative), and those who volunteered to be interviewed contacted the interviewers. The comparison sample was obtained by asking the mothers of children with learning difficulties to name two or three other mothers they knew in their neighbourhood, whose children had not been so labelled, and picking the one whose family composition most nearly resembled that of the mother who nominated them. (As the new city consists of areas which are very different in terms of class and occupational distribution but is reasonably homogeneous *within* areas, asking for local mothers to be nominated yielded a sample matched to a large extent for social class). The interviews were of a loose 'life-history' variety (see Chapter 4 for a discussion of kinds of interviews), covering the women's lives (including paid employment) up to marriage, from marriage to the birth of their children, what changes the children had brought about in their lives, and the history of themselves and the children from birth.

Looking at the mothers' lives, it was evident that the same kind of physical work was involved in bringing up all young children, but the work of the mothers of children with learning difficulties went on for longer and was consequently more intense; if a child remains incontinent until 6 or 7, or even into his or her teens, the bulk of cleaning and washing tasks is very great. The most severely damaged child in the sample was blind, deaf and had no control of limbs; she needed help to sit up or turn over in bed, and washing her hair required two people (one to hold her over the sink while the other did the washing). Some help was received from family and friends (though *less* by the mothers of children with learning difficulties than the other mothers), but six of the 16 mothers in the sample received no help at all, and in five other cases the help was comparatively trivial.

At a broader level of analysis, the families of children with learning difficulties had 'social tasks' to perform over and above those of other parents. Accepting the child at all can be emotional labour:

> For about a month after I found out I didn't have any feeling for her in any way – she wasn't my baby, she was just a baby that had got to be looked after and fed and kept clean. I couldn't pick her up and cuddle her or nothing … And I walked past the pram one day and she looked up at me and she smiled … she just smiled … after that I was all right.

There are social adjustments to be made, and the decision about how to 'put a public face' on the fact of handicap:

> The sooner people knew, I thought, the nicer for them, because there's nothing worse than looking in a pram and it's a friend, and thinking, 'Oh goodness, what can I say?'

Some children fit well into their local communities:

> He chucks his wheelchair around the street and everyone knows him, he goes in next door and has an hour in there and a cup of tea and biscuits, and then he goes off down the road, the old people love him.

But there can be problems, even with immediate family:

> Let's put it this way, there were relations we have not seen since we found out about Trevor … [and] we have only been invited to tea with Trevor once to my brother-in-law. He thinks we should put Trevor away.

Most wearing of all is not knowing, but *thinking*, what attitude other people are taking:

> When I talk to people and I say, 'Mark is mentally handicapped', and as soon as they know he is coming up to sixteen, you see, you know what I mean? I don't want to put it into words, but you see it even before they say it … It is an unspoken look. I suppose maybe I would be guilty in the same way, but there is that fear of danger to 'my daughter'.

The authors summarize the overall results of the study thus:

> having a mentally handicapped child and caring for him or her at home presents a child's parents with two major tasks which are not faced in the same way by parents of 'normal' children. They have to come to terms with the fact of the child's handicap and its implications for the way in which the family is able to conduct its normal life in interaction with others. At the same time they have to deal with the way our society labels and stigmatizes mental handicap – including the way that the historically determined stereotype of mental handicap spills over as a courtesy stigma for the whole family – and this means renegotiating the nature of the family's identity and building a style of life compatible with the renegotiated identity. The task is made none the easier by the fact that the parents are themselves members of the culture which stigmatises them and their children, may project their own feelings of spoilt identity onto the world at large and share to some extent the very attitudes which they are forced to combat. (Abbott and Sapsford, 1987b: 55–6)

It is worth noting that, contrary to expectation, it was not always the child with learning difficulties who was 'the problem' for the family. One mother, for example, was more worried about her eldest child, who had a spell of truanting from school in response to bullying by local children which followed his evident grief at his grandmother's death. Another mother had a child suffering from cystic fibrosis who required daily medical and nursing attention. Another had a child who had been 'teacher's pet' at a small village school and was not adapting to the transfer to a larger secondary school in the new city.

Activity 1.6 (allow 20 minutes)

Look back over the foregoing description in the light of the issues raised so far in this chapter, and make notes on how far it is possible to generalize from the sample and how the nature of the information differs in this study from the ones discussed earlier. Also consider the role of the comparison group in the research.

In terms of representation, this study is clearly inferior to the others already described. The studies we have looked at so far either count all of a class of events or draw a sample which is calculated to be representative in detail of the parent population. In this case we have a sample which cannot be guaranteed to be representative; when the authors say that six of the 16 families did such and such, it would be a mistake to suppose that 37.5 per cent of the parent population would do likewise, and we have no basis for saying how large the error in estimation of percentages actually is. In fact, we know that the sample is biased: it is a volunteer sample – so those who volunteered are likely to have had something they wanted to say – and one of the two schools deliberately excluded certain categories of potential informant (those known to be in distress). The aim was to talk to a typical range of cases, and this may well have been achieved, but we are unable to say the extent to which the sample is unrepresentative of the population.

Beyond this, it is typical of studies of this kind that the data are not precise measurements but rich and complex conversational material. What was picked out of the conversation and included in the analysis was what the authors thought it was important to present to the reader. Even in the full report it would not have been possible to include unedited transcripts (an hour's conversation typically running to 20 pages of transcript). In terms of counts of, for example, amount of help given, the report is undoubtedly factually accurate (within the limitations of decisions about what counts as 'giving help', which is a problem shared by more quantified research). When describing the lives of the mothers, a degree of interpretation necessarily intrudes; although the authors quote a fair amount of what was said to them, this is illustration rather than, strictly, evidence. Where they are framing overall conclusions about the mothers' lives or classifying them into types, what they are doing is to 'give an account' of what they perceive to be true of the data – to 'tell a story' on the basis of the factual evidence.

Furthermore, it is a story based on data which are themselves presented as a story. The informant tells her story, in particular circumstances and to particular interviewers, with a particular perception of what the interviewers are looking to learn and what, therefore, it is relevant to discuss. The interviewers, in their turn, interpret what is said in order to 'tell a story' – or a series of stories of increasing abstraction – based on the material that they have been given. One needs to be well aware of the circumstances of the interview and the predilections of the analysts before what they have to say is acceptable as a plausible account of people's lives.

Against these undoubted weaknesses can be argued the strengths of this kind of approach, all of which revolve around the richness of the data obtained and the comparative lack of structure in the method by which they are collected. One may argue, for example, that even one typical case researched in depth tells us more about a group than superficial information on every member of it: for example, that a thorough exploration of one person's views on religion gives us more insight than figures on church attendance for a whole population. Further, all structured methods involving systematic observation, counting or questionnaires, require decisions *beforehand* as to what is to be relevant, while the relatively unstructured methods used in this study allow surprising information to surface and may question the researchers' preconceptions more readily than more structured approaches.

Pre-test ⟶ Treatment ⟶ Post-test

Figure 1.1 *Diagrammatic representation of simple action research*

Comparing Groups: the Logic of Design

In this section we look at a broader aspect of research as evidence: the logic of the arguments that can be based upon it. Whether the research study is structured to make a particular point from the start or the argument is imposed *post hoc* in the process of writing a report, all research reports embody an argument: this or that is the case, on the basis of the evidence presented. The basic point is that most research depends on *comparison* to establish its conclusions.

We have already looked at the importance of comparisons as tools for drawing conclusions, particularly in the study of mothering children with learning difficulties. In this section we look at how comparison is managed, for what purpose and with what success, in those styles of research which are most explicitly based around comparison of groups or 'conditions': quasi-experimental comparisons and true experiments.

Quasi-experimental Analysis: Road Traffic Fatalities

The classic and most often quoted comparison of naturally occurring groups – capitalizing on changes happening 'in the real world' rather than changes introduced by a researcher – concerns the Connecticut 'crackdown' of the 1950s. This was a drastic public programme in one American state re-inforcing police action to curb excessively fast driving, on which a secondary analysis of published figures was carried out by Donald Campbell (Campbell and Ross, 1968; Campbell, 1969). After an unprecedentedly bad year for traffic fatalities, the governor of Connecticut introduced a programme of administrative orders which made it more certain that drivers who exceeded the speed limit would be caught and, if caught, punished. The net result was a decrease in traffic fatalities within the state of roughly 12 per cent. As it stands, this action has many of the qualities of the simplest kind of *action research*, where researchers/practitioners introduce some kind of change and monitor its effects (except that in this case the change was introduced by the authorities, not the researchers). Figure 1.1 illustrates this kind of research diagrammatically.

Activity 1.7 (10 minutes)

What problems do you see with the conclusion that the governor of Connecticut's action caused fatalities to fall by 12 per cent? Stop and make notes on alternative explanations for the results.

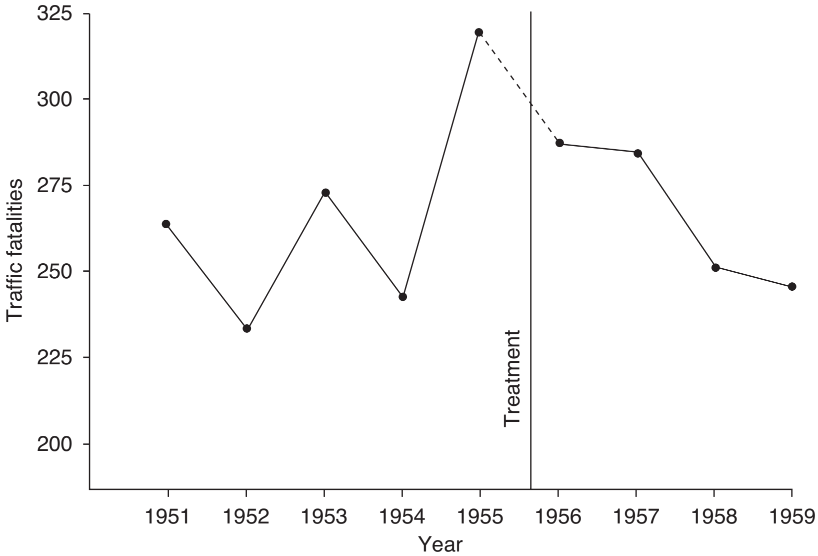

Figure 1.2 *Connecticut traffic fatalities, 1951–1959 (Campbell, 1969: Figure 2)*

Several problems occurred to Campbell, but most of them were answerable by further analysis.

1 There could have been a *trend effect* – the figures could have been going down over the years in any case – coupled with a *regression effect*. (*Regression to the mean* is the technical term for what happens when you look at figures which fluctuate widely around an underlying mean or average trend. Because there is a mean or trend it is very likely, if you have picked an extreme fluctuation, that any other figure will be closer to the mean – and you will remember that the crackdown was initiated because of an *unprecedented* fatality rate.) To explore this we need to compare figures over a longer period than just two years, which is what Campbell did (Figure 1.2). As you can see, there is no obvious trend to explain the decrease in 1956; the figures tend upwards to 1955 and downwards thereafter.
2 There was the possibility that recording practices had changed: for example, by police being less willing to attribute road deaths to speeding. The mere fact of introducing a public programme could have an effect on the record-keeping as well as the driving. Such an effect may well have occurred and the research design is powerless to guard against it.
3 Other equally possible explanations might include: changed weather, changes in petrol prices, changes in drinking behaviour; anything, in fact, that might change driving behaviour. It was possible to guard against most of these, however, by the design of the analysis. What Campbell and his colleagues did was to compare Connecticut with a ‘control group’ – a set of four comparison states in which the crackdown had *not* taken place. Traffic fatality figures from these four states over

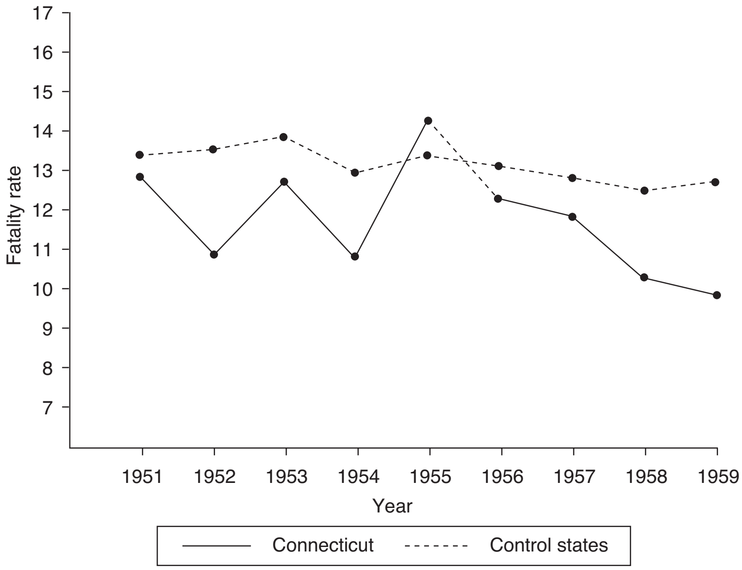

Figure 1.3 *Control series design comparing Connecticut fatalities with those in four comparable states (Campbell, 1969: Figure 11)*

the same period failed to show the downward trend visible in Connecticut (Figure 1.3), and it is therefore likely that none of these factors was responsible for the decrease.

The use of comparison between an 'experimental group' and a 'control group' is a very regular feature of studies which try to show the causal force of treatments, and it is logically very strong. To the extent that the control and experimental groups are alike before treatment, and only the experimental group is treated, if they differ after the treatment the difference must logically be attributable to the treatment. Diagrammatically, this kind of design may be represented by Figure 1.4. The measurements for the two groups will seldom be identical before the treatment, but the real interest is in the difference between pre-test and post-test measurements. In Figure 1.3, for example, the experimental group shows a marked decline from the point of treatment while the control group does not.

Activity 1.8 (5 minutes)

Pause and think: can this design cope with the problem of changes in recording practice listed in Point 2 above?

The use of a control group does not eliminate factors *confounded* with the treatment – things which vary with or as a result of the treatment and therefore cannot be

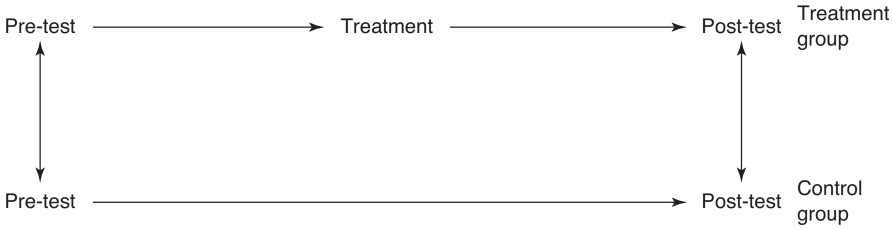

Figure 1.4 *Diagrammatic representation of comparative research using a treatment group and a control group*

distinguished from it. If there were changes in recording practice in Connecticut, for example, these would be specific to the treatment group (they would be part of what happened when the treatment was applied) and could not affect the control group, so the presence of a control group does not eliminate them. What it does make less likely is that a *factor independent of the treatment* was the cause of the decrease in fatalities.

Experiments in HIV Counselling

Consider the following studies:

1. 80 pregnant drug users participating in a methadone maintenance programme were assigned at random to a six-session cognitive/behavioural programme or to a 'control group' who received no such intervention (O'Neill et al., 1996).
2. 119 drug users were randomly assigned to a 50-minute or a 15-minute counselling session (Gibson et al., 1999).
3. 295 drug users participating in a heroin detoxification were randomly assigned to a 50-minute individual counselling session (discussing options and trying to increase condom use) or to receive a package of brochures which covered the same ground (Gibson et al., 1999).
4. 152 cocaine users were randomly assigned to an experimental programme of three two-hour small-group counselling sessions or to an information programme presenting material on video and in print (Malow et al., 1994).
5. 50 people on methadone maintenance were assigned randomly, either to a programme of active presentation of HIV information in small groups, or to receive only brochures (Sorenson et al., 1994).

In all these studies there were pre-intervention measures of AIDS knowledge and self-report measures of condom use.

Activity 1.9 (10 minutes)

Compare these studies with the Connecticut Crackdown and your notes on the Abbott and Sapsford research. What is the function of control/comparison groups in each case, and how effectively does each use this feature? What is distinctively different about the five studies described above?

The five studies all follow the clean lines of the comparative design which we examined above and illustrated in Figure 1.4. One group of subjects received the treatment and the other, picked to be similar to the first, did not, or two treatments are compared; both were tested before and after treatment was applied to one group. The five studies differ from the others described earlier in two important respects:

1 The researchers had control of the treatment. In both the other studies the researchers were looking at 'naturally occurring' variation: in one case, an administrative change to police and court procedure, initiated by a state governor, and, in the other, the natural experience of having and bringing up a child with learning difficulties. In the five studies, on the other hand, it was the researchers who determined what the treatment should be and how it should be run. This clearly gives more opportunity for eliminating possible alternative explanations for the results.
2 The similarity of the groups can to a large extent be guaranteed in the five studies; groups were picked randomly from the same population, which maximizes the chance that anything of importance will be equally distributed between the two groups. In the other studies, the researchers had to 'make do' with naturally occurring dimensions of difference. Campbell and his colleagues had to argue quite carefully and extensively for the validity of the comparison of Connecticut with the four 'comparison states': the extent to which the states used for comparison were really comparable with the state where the 'treatment' occurred. Abbott and Sapsford made some attempt to match the two groups, on social class and family composition, but the success of their efforts is by no means guaranteed.

The difference between the five studies and Campbell's work is that the five studies are *experimental*, while Campbell's is at best a *quasi-experimental* design (to use Campbell's term). An experiment (also called a *controlled trial* in the medical world) is defined as a study in which:

1 The allocation to treatment or control group is under the control of the researcher and can be arranged to maximize the likelihood of having comparable groups.
2 The treatment is under the control of the researcher and can be arranged to minimize the likelihood of other alternative explanatory factors being confounded with its effects.

Activity 1.10 (5 minutes)

What do you see as the advantages of experimental over quasi-experimental designs? Are there corresponding *dis*advantages?

The same logic underlies both kinds of study. The advantage of experimental designs is that the logic is more clearly applied, eliminating a greater number of possible alternative explanations by the design of the study itself. If you measure the state of a group of people at the outset, then administer a treatment, produce measurable changes and *can guarantee that no factor other than the treatment could have produced the effect*, then you are on strong ground in arguing that the treatment produced the effect. Quasi-experimental studies, which capitalize on existing differences between people and/or their circumstances rather than direct manipulation by the experimenter, can never quite offer this guarantee: there are always other potential explanatory factors which might have produced the observed effect. The best you can do is to explore and eliminate those which occurred to you and on which you therefore collected data, by means of statistical control at the analysis stage (see Chapters 9 and 10).

We should note that logical design does not guarantee results which uphold the researcher's prior expectations. In these five studies, for example, the results were:

1 No difference in reported condom use was reported by either group at three months or at 12 months after the study.
2 There was no difference between the groups 9 months after the study.
3 Six- and twelve-month follow-up studies indicated no difference between the groups.
4 Both groups showed a significant reduction in sexual risk-taking three months later, but there was no significant difference between them. However, there was a trend for previously high-risk 'players' on the experimental programme to show greater reduction than those on the information programme.
5 Both groups showed enhanced knowledge immediately after the programme, with those who had attended the active small-group programme knowing significantly more than those who had received only brochures. The difference between the groups was not apparent at three-month follow-up, however.

In other words, none of these studies suggests that counselling or the active provision of information has a lasting effect on subsequent behaviour. It is characteristic of good research design that what the researcher wants to support stands a good chance of *failing* to achieve support; unbiased design provides good support for conclusions, when it does so, precisely *because* the result could have gone the other way.

Four major disadvantages or weaknesses of experimental studies occur to us:

1 It is not possible, even in principle, to *guarantee* that no factor other than the treatment could have produced the effect. You can control for what you know to be important, by the design of the study or by statistical control after the event (see Chapter 9), and you can use randomization techniques to try to even out every other difference between groups, but you can never be sure that you have succeeded. It may still be necessary to control for sources of variation.
2 Some factors cannot be manipulated. Studies of gender, class, age and so on will always have to be at most quasi-experimental because we cannot allocate people to different conditions of them. Similarly, the comparison of risk-takers and others in the fourth study is quasi-experimental; it capitalises on pre-existing difference.
3 Even where the explanatory variable under consideration does lend itself to manipulation and to the allocation of people to one condition or another, ethical considerations may require us to confine ourselves to a fairly small manipulation,

and this may trivialize what we are studying. (The ethics of social research are discussed in more detail in Chapter 15).

4 An added factor about experiments is that they are often very obviously 'research', and this could provide an alternative explanation of any results obtained. If people know they are in a research situation, this in itself can change their behaviour or attitudes. (This is not a *necessary* characteristic of experimental research, however; for example, some experiments are unobtrusive and not obvious to their subjects, while most surveys are obviously a 'research situation', which may well affect the nature of the response).

Activity 1.11 (allow 20 minutes)

The experiment appears to follow a very clear and indisputable form of logic: if we have two identical groups, intervene to administer a treatment to one and not to the other (or administer two different treatments), control all other possible differences and produce a difference in outcome, then the outcome must have been due to the intervention. There are still problems with it, however; spend some time thinking about such research and see if you can identify them.

The weakness of experimental logic, as outlined here, is that it omits to consider part of the context of the argument.

1 In the perfectly designed and executed experiment it may be possible to show *what* caused the effect, but it will not be possible to show *why*. An experiment does not demonstrate the truth of a theory; a variable may have its effect for the reason the researcher posits or for a wide variety of other reasons.
2 The logic of an experiment argues from cause to effect via control of other variables. In many cases, however, the argument that other variables have been controlled, or are not related and may be ignored, will itself rest on a body of theoretical assumptions, and these are not tested by a given experiment but taken for granted. Thus, when an apparent disproof is produced, we shall not know whether the fault is in the theory being tested or in one of the assumptions being taken for granted.
3 Finally, and perhaps most importantly, what the variables mean is not a thing measured but something interpreted on the basis of a whole body of theory. Again, if a negative finding is produced, we shall not know whether the fault lies in the theory under test or in the wider body of theories which define the situation and the meaning of what is being measured.

In these five studies the interpretation is not helped by the inherent difficulty in precise definition of the independent variable (the intervention). All of the papers cited go to considerable lengths to define what *kind* of counselling or cognitive work or information provision is on offer to the experimental group, under what circumstances, how often and for how long. Nonetheless, if there were positive results and they failed to replicate (to be repeated in another precisely similar experiment) it might be because of subtle but important differences in the different applications of supposedly the same counselling or information provision, by different counsellors or teachers, to subtly different populations.

We might go on to question some experimental programmes from another angle – that is, their ethics. It is a perennial ethical problem of research into medical or psychiatric treatments, for example, or into new and efficient ways of schooling, that the efficacy of the treatment can best be demonstrated by applying it to some but withholding it from others. You will note that most of the five studies above do not withhold treatment or information from the control/comparison group but contrast one form or duration of intervention with another. The ethical problems of research which causes pain or distress to its subjects or informants, or in some way disadvantages them, are not confined to experimental research, though they tend to be most obvious there. One may reasonably ask, for example, whether it is justified to cause distress to people who have suffered in the past by opening up the areas of their suffering and pursuing them in interview, simply 'to pursue the truth'. External examiners of academic courses frequently have to ask whether students who explore areas such as AIDS or sexual abuse for their compulsory research dissertations have legitimate access to the area already – for example, they are already doing counselling in the area – or whether they are just trading on human misery in order to do an exciting third-year project. The whole question of the use of people and their experiences as 'research fodder' – the treatment of people as objects by researchers – has been opened up in recent years by, for example, feminist theorists (see Mies, 1993). It is a question to which we shall return in Chapter 15.

Establishing Boundaries in Qualitative Studies

So far we have looked mostly at 'quantitative' studies: research which yields data in the form of numbers to be analyzed by means of comparisons. The logic of comparison also has a large part to play, however, in 'qualitative' studies – ones where the data are in the form of people's words or the researcher's descriptions of what he or she has observed and experienced. For example, in the Abbott and Sapsford (1987b), study which we considered above, if you describe the lives of *mothers who have children with learning difficulties* you are necessarily at the same time describing *mothers with children*; the two are inevitably confounded. It is only by having a group of mothers whose children do *not* have learning difficulties that you can draw more specific conclusions with any degree of validity. What the two groups have in common will be what is true of 'having children', but the areas in which they differ will be aspects specific to 'having children with learning difficulties'. The presence of a comparison group acts to draw a boundary around the conclusions, enabling the researcher to say what is true of the larger group and what holds only for the target group.

This kind of use of a comparison group is a special case of a more general principle of design, which we can illustrate best by means of a fictional example. Suppose we had an interest in why girls tend not to go into science and mathematics at school, except for biology, and to be over-represented in the arts and humanities. We start by 'exploring the field' in a relatively unfocused way, watching and listening to classes, visiting homes, hanging around places where young people go in the evening, and so on. (Already we have used a form of sampling: sampling of the range of contexts in which young people are active and learn and express their ideas). Gathering data, we progressively focus our original vague question down into something more concrete and explorable. We find, let us say, that among working-class girls it is difficult to say what puts them off the sciences and mathematics because everything seems to push in the same direction: teachers, parents, friends and boyfriends all seem to express surprise at or distaste for a girl becoming involved in the sciences. Among middle-class girls, however, we might find that parents and teachers are positively supportive of any interests they might have in the sciences. Their friends, however, and particularly their male friends, seem to treat them as something strange if they opt for the sciences. We have a tentative model of what is going on, then: something in the 'boy/girl culture' is having the effect of making some curriculum choices less attractive than others.

A first stage of drawing boundaries would be to sample classes which were very similar to the ones we used initially – possibly other classes in the same school. If the model fits them as well, then at least it has some generality; if it does not, then there was something idiosyncratic about the particular classes with which we started. Let us say that it does turn out to have some generality. We should then want to sample more widely, to see *how much* generality it has. Does it hold for other classes in the same city? Does it hold for other regions? Does it hold for other English-speaking countries? Each of these comparisons is an attempt to see how widely the idea generalizes: to find the boundaries within which an idea is useful, or the conditions under which a theory holds.

Finally, we might want to start testing particular ideas and assumptions by careful sampling of unlikely milieux. Our tentative model is beginning to be cast in terms of a Western English-speaking 'culture of femininity', let us say. Does it hold, therefore, in schools where the predominant 'culture of femininity' has a potentially different origin? We could test this by finding schools where the majority of students are of Asian origin, say, and seeing whether the same model holds. Above all, our model is about the interaction of the genders, so a very interesting 'crucial case' would be to see what goes on at single-sex schools. If the matter is determined simply by school interaction, then the model should not be as useful in single-sex schools. Suppose, however, that the girls who went in for science and mathematics were seen differently even within single-sex schools. It *could* still be gender-related: there could still be 'masculine' and 'feminine' stereotypes, with the cross-gender roles taken by people of inappropriate gender in single-sex schools where those of the appropriate gender are not available. In that case, we might expect wider stereotyping – that these people acquired inappropriate sex-linked stereotypes outside the field of their academic choices as well. It might be that we were quite mistaken, and gender is not the determining variable that we were inclined to think it was. It could be that the influence of parents or boyfriends outside the school was greater than we had supposed. Whatever the case, there would be further research to be done before we came up with a satisfactory model of what was going on.

In other words, without a carefully constructed basis of comparison we are not able to say precisely what it is that has been found out. We cannot say what is specific to girls in schools without contrasting them with boys, nor whether it is specific to the mixing of girls and boys in classes without contrasting mixed schools with single-sex ones.

Conclusion

We can see that the act of comparison is a central logical device for establishing the validity of a line of argument within research. We use comparison to say why a group is interesting, what about it is interesting, and by how much it differs from expectation.

Planned comparisons are a central element of research design: they are what enables us to draw conclusions and to determine the range over which our conclusions hold true.

Unplanned or *ad hoc* comparisons in the course of a research programme may also shape the initial idea into a firm conclusion and allow it to be put forward as a finding, supported by evidence as well as argument. We know very little about anything except in comparison with something else. In experimental survey research another crucial issue is validity of measurement – the extent to which what is measured and analyzed does actually reflect what the researcher is seeking, and the extent to which the reader can clearly tell that it does so.

In less structured research we do not have 'variables' to 'measure', but it is still important that the circumstances of the data collection assure us that the evidence which is brought forward is plausibly interpreted.

Comparison between groups may be equally as important here as in more 'quantitative' research, in allowing us to establish boundaries for the group about whom our conclusions are true.

Key Terms

Confounded variables/elements aspects of the data which generally cannot be separated even by statistical means (e.g. age and historical period in which born, in a one-shot survey), or biological gender and experience of having been socialized as a male or a female.

Control the imposition of structure on data in order to distinguish the effects of different factors or variables.

- *Design control:* designing data-collection to ensure that no variables are *confounded* (see above) – as in the experiment.
- *Statistical control*: distinguishing the effects of variables at the analysis stage, as in most survey analysis.

Constructionism a view of the social world as a product of history and developing structures, and of human behaviour as societally and historically more than biologically defined.

Experiment a study in which precisely measured interventions are applied to groups selected by the experimenter (randomly or by matching characteristics) as identical at the start of the research, so that any measured outcomes can unambiguously be attributed to the intervention.

Holism treating the research situation as a whole – the opposite of *reductionism* (below).

Interactionism a view of the social world as a product of the interacting meaning-systems, and actions of people and groups and of human behaviour as socially more than biologically defined.

Longitudinal (cohort) survey designs ones which take measurements, repeated over time, from the same people or units.

Naturalism trying to disturb the natural situation as little as possible; acknowledging that the research situation modifies the natural situation and therefore trying to minimize the amount of imposed structure.

One-shot surveys surveys which collect data at one point of time only (often involving retrospective measurement of what has occurred in the past).

Positivism the view that social scientific research should follow the principles and methods of the physical and biological sciences, that the major problems are problems of measurement and that the researcher can and should remain external to the research.

Quasi-experimental analysis analysis of survey data which tries to follow the logic of *experimental* research (see above) but were collected from naturally occurring rather than experimenter-selected groups.

Reductionism identifying the logical elements of a situation or process, for separate and uncontaminated study.

Regression to the mean if there is an underlying trend in figures, with data-points varying randomly around it, and if you select a data-point reasonably remote from the mean, then it is highly probably that the next data-point will be closer to it.

Time-series (trend) survey designs studies which take repeated measurements over time but draw a fresh sample each time.

Validity the extent to which the research conclusions can plausibly be taken to represent a state of affairs in the wider world.

- *Population validity:* the extent to which a sample may be taken as representing or typical of the population from which it is drawn.
- *Validity of measurement:* the extent to which we are assured that the measurements in the research do indeed represent what the researcher says they represent and are not produced by the research process itself.

Further Introductory Reading

Jupp, V. (1989) *Methods of Criminological Research*, London, Allen and Unwin (Chapter 2).
Sapsford, R.J. (1999) *Survey Research*, London, Sage (Chapters 1 and 2).
Sapsford, R.J. and Abbott, P.A. (1998) *Research Method for Nurses and the Caring Professions*, 2nd edn., Buckingham, Open University Press (Chapter 1).

Summary Activity: Preparing Your Research Proposal

One of the key aims of this book is to develop the ability to read research reports critically, in a way that facilitates the assessment of validity. A second aim is to develop abilities in the production of research ideas and, more specifically, research proposals and strategies for examining such ideas (for example, proposals for thesis/dissertation research). The aim is to be able to plan research that will produce conclusions that other readers will find valid and credible. For this reason, you will find an activity at the end of each chapter which relates to the production of a research proposal. Each time the activity recurs it will provide a set of questions or prompts relating to material in the chapter. Not all of these questions will be relevant to your own proposed research, but they are all worthy of being addressed – if only to rule them out after assuring yourself that the issues *have* been addressed.

You add to your proposal as you progress through each chapter and as you move from considering design issues through to data collection, data analysis and the drawing of conclusions from such analysis. Cutting across these stages are questions concerning the choice of quantitative as opposed to qualitative data, the choice of primary data collection as opposed to secondary analysis, and questions about the ethics and politics of your research. The first version of the activity draws on material from this chapter and asks you to address questions such as 'Who should I include in my design?' (case selection), 'What data should I collect about them?' (data collection) and 'Over what period of time?' (comparison over time).

1 This chapter has emphasized the importance of validity: that is, designing research such that possible conclusions about the research problem can flow logically from the evidence generated. Valid answers start with clear questions, so what is the research problem or question at the centre of your proposed research?
2 Given the aim of seeking to reach valid conclusions about the research problem, what form of research design (or combination of these) would seem appropriate?

   - one based on secondary analysis of data?
   - survey-based research?
   - naturalistic or qualitative research?
   - quasi-experimental analysis?
   - an experimental design?

3 With regard to case selection:

   - what cases should be selected (individuals, groups of individuals, interactions, social settings, etc.)?
   - how should these be selected (at random, using volunteers, taking naturally occurring groups, events or settings, etc.)?
   - are the cases to be selected typical or even representative of the population of individuals, groups or contexts about which you wish to draw conclusions?

4 (a) In what ways does your case selection facilitate comparison between individuals, groups or settings (for example, comparison of the answers of different groups of individuals to the same questions, comparison of the behaviour of the same people in different settings, comparison of figures from different years, comparison of present experience and remembered experience in the past, comparison of documents produced by differently 'situated' sources)?

   (b) Are these comparisons appropriate to the research problem you are addressing, and will they be able to yield valid conclusions about the problem?

5 In what ways does your case selection allow comparison over time, and how does this relate to the research problem in terms of, say, examining changes in society (trend studies) or changes in individuals (cohort studies)? Is comparison over time at all relevant to your research question?
6 What measurements, if any, does your research problem require, and how will these be collected (for example, at first hand or by accessing secondary data)? Is it likely that the figures produced will be justifiable as measurements of what we want them to measure (validity of measurement).