

Research methods: Data analysis

- **Qualitative analysis of data**
Recording experiences and meanings
Distinctions between quantitative and qualitative studies
Reason and Rowan's views
- **Interpretations of interviews, case studies, and observations**
Some of the problems involved in drawing conclusions from non-experimental studies.
Reicher and Potter's St Paul's riot study
McAdams' definition of psychobiography
Weiskrantz's study of DB
Jourard's cross-cultural studies
- **Content analysis**
Studying the messages contained in media and communications.
Cumberbatch's TV advertising study
A bulimia sufferer's diary
- **Quantitative analysis: Descriptive statistics**
What to do with all those numbers and percentages at the end of the study.
Measures of central tendency: Mean, median, and mode
Levels of measurement
Measures of dispersion: range, interquartile range, variation ratio, standard deviation
- **Data presentation and statistical tests**
When to use a chart or a graph. Which statistical test to choose and why.
Frequency polygon, histogram, and bar chart
Types of data: nominal, ordinal, interval, ratio
Statistical significance
Tests of difference: Mann-Whitney U test, sign test, Wilcoxon test
Scattergraphs, Spearman's rho
Test of association: chi-squared
- **Issues of experimental and ecological validity**
Does your study test what you say it does? Has it any relevance to real life?
Questions to test experimental validity
Varied definitions of ecological validity
- **Writing up a practical**
Presenting your results.
Style and approach
Headings to use

Note: Cross references to "PIP" refer to the printed version of Psychology: An International Perspective by Michael W. Eysenck.

The data obtained from a study may or may not be in numerical or quantitative form, that is, in the form of numbers. If they are not in numerical form, then we can still carry out qualitative analyses based on the experiences of the individual participants. If they are in numerical form, then we typically start by working out some descriptive statistics to summarise the pattern of findings. These descriptive statistics include measures of central tendency within a sample (e.g. mean) and measures of the spread of scores within a sample (e.g. range). Another useful way of summarising the findings is by means of graphs and figures. Several such ways of summarising the data are discussed later on in this chapter.

In any study, two things might be true: (1) there is a difference (the experimental hypothesis), or (2) there is no difference (the null hypothesis). Various statistical tests have been devised to permit a decision between the experimental and null hypotheses on the basis of the data. Decision making based on a statistical test is open to error, in that we can never be sure whether we have made the correct decision. However, certain standard procedures are generally followed, and these are discussed in this chapter.

Finally, there are important issues relating to the validity of the findings obtained from a study. One reason why the validity of the findings may be limited is that the study itself was not carried out in a properly controlled and scientific fashion. Another reason why the findings may be partially lacking in validity is that they cannot readily be applied to everyday life, a state of affairs that occurs most often with laboratory studies. Issues relating to these two kinds of validity are discussed towards the end of the chapter.

How would you define "validity"? How does it differ from "reliability"?

QUALITATIVE ANALYSIS OF DATA

There is an important distinction between quantitative research and qualitative research. In quantitative research, the information obtained from the participants is expressed in numerical form. Studies in which we record the number of items recalled, reaction times, or the number of aggressive acts are all examples of quantitative research. In qualitative research, on the other hand, the information obtained from participants is *not* expressed in numerical form. The emphasis is on the stated experiences of the participants and on the stated meanings they attach to themselves, to other people, and to their environment. Those carrying out qualitative research sometimes make use of direct quotations from their participants, arguing that such quotations are often very revealing.

Quantitative research would measure the number of aggressive acts witnessed. Qualitative research may help to explain why the aggressive acts occurred.

There has been rapid growth in the use of qualitative methods since the mid-1980s. This is due in part to increased dissatisfaction with the quantitative or scientific approach that has dominated psychology for the past 100 years. Coolican (1994) discussed a quotation from Reason and Rowan (1981), which expresses that dissatisfaction very clearly:

There is too much measurement going on. Some things which are numerically precise are not true; and some things which are not numerical are true. Orthodox research produces results which are statistically significant but humanly insignificant; in human inquiry it is much better to be deeply interesting than accurately boring.

Many experimental psychologists would regard this statement as being clearly an exaggeration. "Orthodox research" with its use of the experimental method has transformed our understanding of attention, perception, learning, memory, reasoning, and so on. However, qualitative research is of clear usefulness within some areas of social psychology, and it can shed much light on the motivations and values of individuals. As a result, investigators using interviews, case studies, or observations often make use of qualitative data, although they do not always do so.

Investigators who collect qualitative data use several different kinds of analysis, and so only general indications of what can be done with such data will be presented here. However, there would be general agreement among such investigators with the following statement by Patton (1980; cited in Coolican, 1994):

The cardinal principle of qualitative analysis is that causal relationships and theoretical statements be clearly emergent from and grounded in the

phenomena studied. The theory emerges from the data; it is not imposed on the data.

How do investigators use this principle? One important way is by considering fully the categories spontaneously used by the participants *before* the investigators develop their own categories. An investigator first of all gathers together all the information obtained from the participants. This stage is not always entirely straightforward. For example, if we simply transcribe tape recordings of what our participants have said, we may be losing valuable information. Details about which words are emphasised, where the speaker pauses, and when the speaker speeds up or slows down should also be recorded, so that we can understand fully what he or she is trying to communicate.

The investigator then arranges the items of information (e.g. statements) into various groups in a preliminary way. If a given item seems of relevance to several groups, then it is included in all of them. Frequently, the next step is to take account of the categories or groupings suggested by the participants themselves. The final step is for the investigator to form a set of categories based on the information obtained from the previous steps. However, the investigator is likely to change some of the categories if additional information comes to light.

Qualitative investigators are not only interested in the number of items or statements falling into each category. Their major concern is usually in the variety of meanings, attitudes, and interpretations found within each category. For example, an investigator might study attitudes towards A-level psychology by carrying out interviews with several A-level students. One of the categories into which their statements were then placed might be “negative attitudes towards statistics”. A consideration of the various statements in this category might reveal numerous reasons why A-level psychology students dislike statistics!

When qualitative researchers report their findings, they will often include some raw data (e.g. direct quotations from participants) as well as analyses of the data based on categories. In addition, they often indicate how their hypotheses changed during the course of the investigation.

How might your own learned cultural experiences determine how you view others' behaviour?

Investigators must take care that cultural bias does not lead to their own values, norms, and beliefs distorting the data they collect.

Evaluation

Qualitative analysis is often less influenced than is quantitative analysis by the biases and theoretical assumptions of the investigator. In addition, it offers the prospect of understanding the participants in a study as rounded individuals in a social context. This contrasts with quantitative analysis, in which the focus is often on rather narrow aspects of behaviour.

The greatest limitation of the qualitative approach is that the findings that are reported tend to be unreliable and hard to replicate. Why is this so? The qualitative approach is subjective and impressionistic, and so the ways in which the information is categorised and then interpreted often differ considerably from one investigator to another.

There are various ways in which qualitative researchers try to show that their findings are reliable (Coolican, 1994). Probably the most satisfactory approach is to see whether the findings obtained from a qualitative analysis can be replicated. This can be done by comparing the findings from an interview study with those from an observational study. Alternatively, two different qualitative researchers can conduct independent analyses of the same qualitative data, and then compare their findings.

Qualitative researchers argue that the fact that they typically go through the “research cycle” more than once helps to increase reliability. Thus, for example, the initial assumptions and categories of the researcher are checked against the data, and may then be changed. After that, the new assumptions and categories are checked against the data. Repeating the research cycle is of value in some ways, but it does not ensure that the findings will have high reliability.

■ Research activity: Categorising television programmes

In small groups of three or four people, consider how you might conduct a study to analyse the number of aggressive acts witnessed by children when they watch television cartoon programmes designed for a child audience. Would quantitative methods be most appropriate? How would you ensure that your results were as reliable as possible?

INTERPRETATION OF INTERVIEWS, CASE STUDIES, AND OBSERVATIONS

Qualitative analyses as discussed in the previous section are carried out in several different kinds of studies. They are especially common in interviews, case studies, and observational studies, although quantitative analyses have often been used in all three types of studies. Some of the advantages and limitations of these types of studies are discussed in the Research methods: Psychological enquiry chapter. What we will do in this section is to consider the interpretation of interviews, case studies, and observations.

Interviews

As discussed in the Research methods: Psychological enquiry chapter, interviews vary considerably in terms of their degree of structure. In general terms, unstructured interviews (e.g. non-directive or informal) lend themselves to qualitative analyses, whereas structured interviews lend themselves to quantitative analysis. As Coolican (1994) pointed out, there are various skills that interviewers need in order to obtain valuable data. These skills involve establishing a good understanding with the person being interviewed, adopting a non-judgemental approach, and developing effective listening skills.

Cardwell et al. (1996) illustrated the value of the interview approach by discussing the work of Reicher and Potter (1985) on a riot in the St Paul's area of Bristol in April 1980. Many of the media reports on the riot were based on the assumption that those involved in the riot were behaving in a primitive and excessively emotional way. Unstructured interviews with many of those involved indicated that in fact they had good reasons for their actions. They argued that they were defending their area against the police, and they experienced strong feelings of solidarity and community spirit. This interpretation was supported by the fact that very little of the damage affected private homes in the area.

Evaluation

There are various problems involved in interpreting interview information.

First, there is the problem of **social desirability bias**. Most people want to present themselves in the best possible light, so they may provide socially desirable rather than honest answers to personal questions. This problem can be handled by the interviewer asking additional questions to establish the truth.

Second, the data obtained from an interviewer may reveal more about the social interaction processes between the interviewer and the person being interviewed (the interviewee) than about the interviewee's thought processes and attitudes.

Third, account needs to be taken of the **self-fulfilling prophecy**. This is the tendency for someone's expectations about another person to lead to the fulfilment of those expectations. For example, suppose that a therapist expects his or her patient to behave very anxiously. This expectation may cause the therapist to treat the patient in such a way that the patient starts to behave in the expected fashion.

KEY TERMS

Social desirability bias: the tendency to provide socially desirable rather than honest answers on questionnaires and in interviews.

Self-fulfilling prophecy: the tendency for someone's expectations about another person to lead to the fulfilment of those expectations.

Psychobiography: the study of individual personality by applying psychological theory to the key events in a person's life.

Case studies

Case studies (intensive investigations of individuals) come in all shapes and sizes. Probably the best-known case studies are those of Freud and others in the field of clinical psychology. However, detailed case studies have also been carried out in personality research and in studies of cognitive functioning in brain-damaged patients.

One way in which case studies have been used to study personality involves an approach known as **psychobiography**. This was defined by McAdams (1988, p. 2) as "the systematic use of psychological (especially personality) theory to transform a life into a coherent and illuminating story." A key feature of psychobiography is identification of the most important events in an individual's account of his or her own life story. How can this be done? According to McAdams (1988, pp. 12–13), we should look for

CASE STUDY: *The Effects of Extreme Deprivation*

Freud and Dann (1951) studied six preschool children who had lost their parents during the Second World War. It is not known how long each child had spent with their parents before being taken to Nazi concentration camp nurseries. The children remained together, despite moving camp several times, and appeared to have received only the most basic forms of care and attention. In the absence of a caring adult, they had formed close and loving bonds with each other. These strong bonds provided a protective and stable influence in their lives.

The children were rescued at the end of the war and brought to England for medical and psychological treatment. Their mental and physical development had been restricted, so that they had very poor speech skills. They feared adults and clung to each other for reassurance. Gradually they began to form bonds with the adults who cared for them, and their social and language skills improved.

Despite all the problems they had experienced, the children did not show the levels of extreme disturbance that were once expected when there is a complete lack of “mothering” (Bowlby, 1951). Freud and Dann’s study highlights the fundamental importance of having someone to bond with, even if it is not the mother, as well as the reversibility of the effects of extreme deprivation.

Case studies are often seen as rather unscientific and unreliable. The sample is not representative of the wider population, the study cannot be repeated, and interpretation of the findings is very subjective. However, case studies can be of great interest because they highlight unique and unexpected behaviour, and can stimulate research that may contradict established theories such as Bowlby’s. Freud and Dann’s work offers insights into human experience that would otherwise be impossible to gain: ethical considerations prevent the deliberate separation of children and parents in order to study the effects of deprivation.

clues about primacy (what comes first in a story), uniqueness (what stands out in the story), omission (what seems to be missing from the story), distortion and isolation (what doesn’t follow logically in the story), and incompleteness (when the story fails to end in a satisfying way).

Weiskrantz (1986) reported a very different kind of case study. He studied DB, who had had an operation designed to reduce the number of severe migraines from which he suffered. As a result of this operation, DB exhibited what is known as “blindsight”. He was able to tell whether a visual stimulus had been presented, and he could point at it, even though he had no conscious awareness of having seen it. These findings are important, because they suggest that many perceptual processes can occur in the absence of conscious awareness.

Evaluation

We need to be very careful when interpreting the evidence from a case study. The greatest danger is that very general conclusions may be drawn on the basis of a single atypical individual. For this reason, it is important to have supporting evidence from other sources before drawing such conclusions.

It is often hard to interpret the evidence from case studies. For example, Freud claimed that the various case studies he reported served to show the validity of his theoretical ideas. However, such evidence is suspect, because there was a real chance of contamination in the data Freud obtained from his patients. What any patient said to Freud may have been influenced by what Freud had said to him or her previously, and Freud may have used his theoretical views to interpret what the patient said in ways that distorted it.

How, then, should the findings from a case study be interpreted? Probably the greatest value of a case study is that it can suggest hypotheses which can then be tested under more controlled conditions with larger numbers of participants. In other words, case studies usually provide suggestive rather than definitive evidence. In addition, case studies can indicate that there are limitations in current theories. The discovery of blindsight in DB suggested that visual perception depends much less on conscious awareness than was thought to be the case by most theorists.

Observations

As discussed in the Research methods: Psychological enquiry chapter, there are numerous kinds of observational studies, and the data obtained may be either quantitative or qualitative. We will consider issues relating to interpreting the data from observational studies by focusing on a concrete example. Jourard (1966) watched pairs of people talking in cafes, and noted down the number of times one person touched another at one table

during one hour. In San Juan, the capital of Puerto Rico, the total number of touches was 180. In contrast, the total in Paris was 110, and in London it was 0. One problem with interpreting these data is that the kinds of people who go to cafes in San Juan, Paris, and London may be quite different. It is also entirely possible that those who spend much of their time in cafes are not representative of the general population. These issues of representativeness apply to many observational studies.

Evaluation

Jourard's (1966) findings do not really tell us why there is (or was, in 1966) much more touching in San Juan than in London. It is possible that Londoners are simply less friendly and open, but there are several other possibilities (e.g. Londoners are more likely to go to cafes with business colleagues). The general issue here is that it is often very hard to interpret or make sense of the data obtained from observational studies, because we can only speculate on the reasons why the participants are behaving in the ways that we observe.

Another issue was raised by Coolican (1994) in his discussion of the work of Whyte (1943). Whyte joined an Italian street gang in Chicago, and became a participant observer. The problem he encountered in interpreting his observations was that his presence in the gang influenced their behaviour. A member of the gang expressed this point as follows: "You've slowed me down plenty since you've been down here. Now, when I do something, I have to think what Bill Whyte would want me to know about it and how I can explain it."

CONTENT ANALYSIS

Content analysis is used when originally qualitative information is reduced to numerical terms. **Content analysis** started off as a method for analysing messages in the media, including articles published in newspapers, speeches made by politicians on radio and television, various forms of propaganda, and health records. More recently, the method of content analysis has been applied more widely to almost any form of communication. As Coolican (1994, p. 108) pointed out:

The communications concerned were originally those already published, but some researchers conduct content analysis on materials which they ask people to produce, such as essays, answers to interview questions, diaries, and verbal protocols [detailed records].

One of the types of communication that has often been studied by content analysis is television advertising. For example, McArthur and Resko (1975) carried out a content analysis of American television commercials. They found that 70% of the men in these commercials were shown as experts who knew a lot about the products being sold. In contrast, 86% of the women in the commercials were shown only as product users. There was another interesting gender difference: men who used the products were typically promised improved social and career prospects, whereas women were promised that their family would like them more.

More recent studies of American television commercials (e.g. Brett & Cantor, 1988) indicate that the differences in the ways in which men and women are presented have been reduced. However, it remains the case that the men are far more likely than women to be presented as the product expert.

The first stage in content analysis is that of sampling, or deciding what to select from what may be an enormous amount of material. For example, when Cumberbatch (1990) carried out a study on over 500 advertisements shown on British television, there were two television channels showing advertisements. Between them, these two channels were broadcasting for about 15,000 hours a year, and showing over 250,000 advertisements. Accordingly, Cumberbatch decided to select only a sample of advertisements taken from prime-time television over a two-week period.

KEY TERM

Content analysis: a method involving the detailed study of, for example, the output of the media, speeches, and literature.

The issue of sampling is an important one. For example, television advertisers target their advertisements at particular sections of the population, and so arrange for the advertisements to be shown when the relevant groups are most likely to be watching television. As a result, advertisements for beer are more likely to be shown during a football match than a programme about fashion. By focusing on prime-time television, Cumberbatch (1990) tried to ensure that he was studying advertisements designed to have general appeal.

The other key ingredient in content analysis is the construction of the **coding units** into which the information is to be categorised. In order to form appropriate coding units, the researcher needs to have considerable knowledge of the kinds of material to be used in the content analysis. He or she also needs to have one or more clear hypotheses, because the selection of coding units must be such as to permit these hypotheses to be tested effectively.

The coding can take many forms. The categories used can be very specific (e.g. use of a given word) or general (e.g. theme of the communication). Instead of using categories, the coders may be asked to provide *ratings*. For example, the apparent expertise of those appearing in television advertisements might be rated on a 7-point scale. Another form of coding involves *ranking* items, or putting them in order. For example, the statements of politicians could be ranked in terms of the extent to which they agreed with the facts.

Gender and advertising

Cumberbatch (1990) found that men outnumbered women in advertisements by 2:1. In addition, 75% of the men in ads were aged over 30, whereas 75% of women in ads were aged under 30. Male voices were used where the information in the soundtrack concerned technical expertise, whereas women's voices were used in sexy and sensuous ways. What does this say about the way we view men and women in society? Comparing the results of studies such as Cumberbatch's with earlier ones (e.g. McArthur & Resko, 1975) can begin to provide answers to questions such as this.

Evaluation

One of the greatest strengths of content analysis is that it provides a way of extracting information from a wealth of real-world settings. The media influence the ways we think and feel about issues, and so it is important to analyse media communications in detail. Content analysis can reveal issues of concern. For example, Cumberbatch (1990) found in his study of advertisements on British television that only about 25% of the women appearing in these advertisements seemed to be over 30 years old, compared to about 75% of the men. On the face of it, this would seem to reflect a sexist bias.

KEY TERM

Coding units: the categories into which observations are placed prior to analysis.

Food Diary – Week 1

Time	What eaten	B	V	L	Antecedents & Consequences
8:00	All-bran				A: Still full from yesterday. C: Must make an effort not to binge today.
12:00	1 apple				A: Hungry. C: Still hungry, mustn't eat more in case it starts me off on a binge.
3:00	1 lb grapes, 2 choc. bars		!		A: Had phone call from John, he will be home late. C: Disgusted with myself. I am the most hopeless person in the world.
6:00	peanuts + choc, picked from shopping	!!			A: No food in flat. Had to go shopping. Couldn't stop myself putting loads of sweets in the trolley. Ate loads of stuff in the car. Had to go on eating once at home.
7:00	2 portions of corny, 3 choc. bars	!!		!!	C: Very angry with myself. I feel so lonely. Totally exhausted, went to bed early.

B = Binge, V = Vomited, L = Laxatives

Food Diary – Week 4

Time	What eaten	B	V	L	Antecedents & Consequences
8:00	Cottage cheese, 2 sl. toast with honey				Enjoyed this.
11:00	apple				
12:30	baked potato, tuna fish				Eaten in the canteen at work. Tina said "You haven't been here for ages". Could have run away, felt everybody was looking at me.
3:00	yoghurt, crunch bar				
6:00	1 sl. toast				
7:00	fish + vegetables, 1 portion ice cream				Had not planned dessert. John suggested ice cream. My initial response was to say no, but I knew I would then finish the packet off whilst washing up. So I had a portion and enjoyed it sitting with John. John put it away and made coffee, which we drank relaxing on the sofa. Washing up left.

Diary studies are often used in clinical psychology, such as in this example from the diary of a bulimia sufferer. Diaries may be used to record actions, thoughts, and feelings, but may not be totally accurate, particularly if the diarist is embarrassed to reveal the truth about himself or herself. Food diaries reproduced from U. Schmidt and J. Treasure (1993), with permission.

Coding units might include time, space, words, themes, roles, items, actions, etc. How would you use these to analyse the content of television programme such as "soaps"?

The greatest limitation of content analysis is that it is often very hard to interpret the findings. Consider, for example, the difference in the ages of men and women appearing in advertisements found by Cumberbatch (1990). One interpretation is that this difference occurred because most television viewers prefer to see older men and younger women in advertisements. However, it is also possible that those making the advertisements thought mistakenly that this is what the viewers wanted to see. There are other possible interpretations, but the available data do not allow us to discriminate among them.

There are also problems of interpretation with other communications such as personal diaries or essays. Diaries or essays may contain accurate accounts of what an individual does, thinks, and feels. On the other hand, individuals may provide deliberately distorted accounts in order to protect their self-esteem, to make it appear that their lives are more exciting than is actually the case, and so on.

Another problem is that the selection and scoring of coding units can be rather subjective. The coding categories that are used need to reflect accurately the content of the communication, and each of the categories must be defined as precisely as possible.

QUANTITATIVE ANALYSIS: DESCRIPTIVE STATISTICS

Suppose that we have carried out an experiment on the effects of noise on learning with three groups of nine participants each. One group was exposed to very loud noise, another group to moderately loud noise, and the third group was not exposed to noise at all. What they had learned from a book chapter was assessed by giving them a set of questions, producing a score between 0 and 20.

What is to be done with the raw scores? There are two key types of measures that can be taken whenever we have a set of scores from participants in a given condition. First, there are measures of central tendency, which provide some indication of the size of average or typical scores. Second, there are measures of dispersion, which indicate the extent to which the scores cluster around the average or are spread out. Various measures of central tendency and of dispersion are considered next.

Measures of central tendency

Measures of central tendency describe how the data cluster together around a central point. There are three main measures of central tendency: the mean; the median; and the mode.

Mean

The **mean** in each group or condition is calculated by adding up all the scores in a given condition, and then dividing by the number of participants in that condition. Suppose that the scores of the nine participants in the no-noise condition are as follows: 1, 2, 4, 5, 7, 9, 9, 9, 17. The mean is given by the total, which is 63, divided by the number of participants, which is 9. Thus, the mean is 7.

The main advantage of the mean is the fact that it takes all the scores into account. This generally makes it a sensitive measure of central tendency, especially if the scores resemble the **normal distribution**, which is a bell-shaped distribution in which most scores cluster fairly close to the mean. However, the mean can be very misleading if the distribution differs markedly from the normal and there are one or two extreme scores in one direction. Suppose that eight people complete one lap of a track in go-karts. For seven of them, the times taken (in seconds) are as follows: 25, 28, 29, 29, 34, 36, and 42. The eighth person's go-kart breaks down, and so the driver has to push it around the track. This person takes 288 seconds to complete the lap. This produces an overall mean of 64 seconds. This is clearly misleading, because no-one else took even close to 64 seconds to complete one lap.

Median

Another way of describing the general level of performance in each condition is known as the **median**. If there is an odd number of scores, then the median is simply the middle score,

Mean

Scores	Number	
1	1	
2	2	
4	3	
5	4	
7	5	
9	6	
9	7	
9	8	
17	9	
<hr/>		
63	9	Total
63	÷ 9	= 7

KEY TERMS

Mean: an average worked out by dividing the total of all participants' scores by the number of participants.

Normal distribution: a bell-shaped distribution in which most scores cluster fairly close to the mean.

Median: the middle score out of all participants' scores in a given condition.

having an equal number of scores higher and lower than it. In the example with nine scores in the no-noise condition (1, 2, 4, 5, 7, 9, 9, 9, 17), the median is 7. Matters are slightly more complex if there is an even number of scores. In that case, we work out the mean of the two central values. For example, suppose that we have the following scores in size order: 2, 5, 5, 7, 8, 9. The two central values are 5 and 7, and so the median is

$$\frac{5+7}{2} = 6$$

The main advantage of the median is that it is unaffected by a few extreme scores, because it focuses only on scores in the middle of the distribution. It also has the advantage that it tends to be easier than the mean to work out. The main limitation of the median is that it ignores most of the scores, and so it is often less sensitive than the mean. In addition, it is not always representative of the scores obtained, especially if there are only a few scores.

Mode

The final measure of central tendency is the **mode**. This is simply the most frequently occurring score. In the example of the nine scores in the no-noise condition, this is 9. The main advantages of the mode are that it is unaffected by one or two extreme scores, and that it is the easiest measure of central tendency to work out. In addition, it can still be worked out even when some of the extreme scores are not known. However, its limitations generally outweigh these advantages. The greatest limitation is that the mode tends to be unreliable. For example, suppose we have the following scores: 4, 4, 6, 7, 8, 8, 12, 12, 12. The mode of these scores is 12. If just one score changed (a 12 becoming a 4), the mode would change to 4! Another limitation is that information about the exact values of the scores obtained is ignored in working out the mode. This makes it a less sensitive measure than the mean. A final limitation is that it is possible for there to be more than one mode.

The mode is useful where other measures of central tendency are meaningless, for example when calculating the number of children in the average family. It would be unusual to have 0.4 or 0.6 of a child!

Levels of measurement

From what has been said so far, we have seen that the mean is the most generally useful measure of central tendency, whereas the mode is the least useful. However, we need to take account of the level of measurement when deciding which measure of central tendency to use (the various levels are discussed further on p. 15 of this chapter). At the interval and ratio levels of measurement, each added unit represents an equal increase. For example, someone who hits a target four times out of ten has done twice as well as someone who hits it twice out of ten. Below this is the ordinal level of measurement, in which we can only order, or rank, the scores from highest to lowest. At the lowest level, there is the nominal level, in which the scores consist of the numbers of participants falling into various categories. The mean should only be used when the scores are at the interval level of measurement. The median can be used when the data are at the interval or ordinal level. The mode can be used when the data are at any of the three levels. It is the only one of the three measures of central tendency that can be used with nominal data.

Measures of dispersion

The mean, median, and mode are all measures of central tendency. It is also useful to work out what are known as measures of dispersion, such as the range, interquartile range, variation ratio, and standard deviation. These measures indicate whether the scores in a given condition are similar to each other or whether they are spread out.

Range

The simplest of these measures is the **range**, which can be defined as the difference between the highest and the lowest score in any condition. In the case of the no-noise group (1, 2, 4, 5, 7, 9, 9, 9, 17), the range is $17 - 1 = 16$.

Scores

1
2
4
5
9
9
9
17

7 = Median

Scores

1
2
4
5
7
9
9
17

9 = Mode

KEY TERMS

Mode: the most frequently occurring score among the participants in a given condition.

Range: the difference between the highest and lowest score in any condition.

spread or dispersion of the scores. It has the disadvantage that it ignores information from the top and the bottom 25% of scores. For example, we could have two sets of scores with the same interquartile range, but with more extreme scores in one set than in the other. The difference in spread or dispersion between the two sets of scores would not be detected by the interquartile range.

Variation ratio

Another simple measure of dispersal is the **variation ratio**. This can be used when the mode is the chosen measure of central tendency. The variation ratio is defined simply as the proportion of the scores obtained which are not at the modal value (i.e. the value of the mode). The variation ratio for the no-noise condition discussed earlier (scores of 1, 2, 4, 5, 7, 9, 9, 9, 17), where the mode is 9, is as follows:

$$\frac{\text{number of non-modal scores}}{\text{total number of scores}} = \frac{6}{9} = 0.67$$

The advantages of the variation ratio are that it is not affected by extreme values, and that it is very easy to calculate. However, it is a very limited measure of dispersal, because it ignores most of the data. In particular, it takes no account of whether the non-modal scores are close to, or far removed from, the modal value. Thus, the variation ratio can only provide a very approximate measure of dispersal.

Standard deviation

The most generally useful measure of dispersion is the **standard deviation**. It is harder to calculate than the range or variation ratio, but generally provides a more accurate measure of the spread of scores. However, you will be pleased to learn that many calculators allow the standard deviation to be worked out rapidly and effortlessly, as in the worked example.

Standard deviation: A worked example

Participant	Score X	Mean M	Score – Mean X – M	(Score – Mean) ² (X – M) ²
1	13	10	3	9
2	6	10	–4	16
3	10	10	0	0
4	15	10	5	25
5	10	10	0	0
6	15	10	5	25
7	5	10	–5	25
8	9	10	–1	1
9	10	10	0	0
10	13	10	3	9
11	6	10	–4	16
12	11	10	1	1
13	7	10	–3	9
13	130	10		136

Total of scores = $\sum X = 130$

Number of participants = $N = 13$

$$\text{Mean} = \frac{\sum X}{N} = \frac{130}{13} = 10$$

$$\text{Variance} = \frac{136}{13 - 1} = 11.33$$

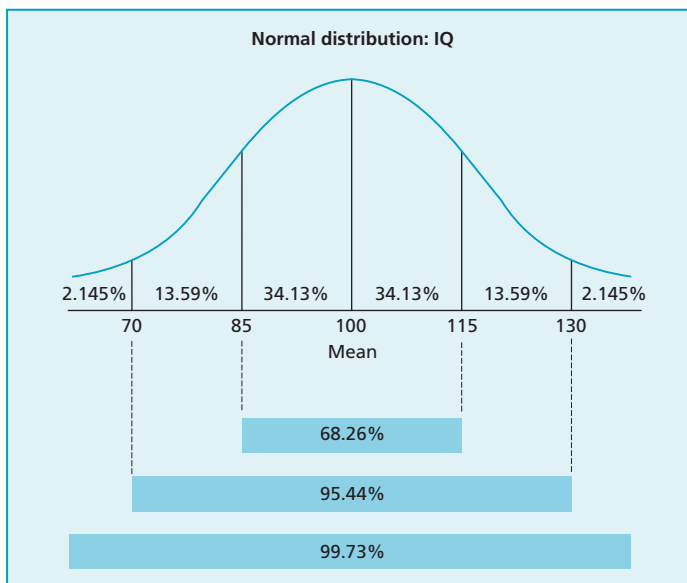
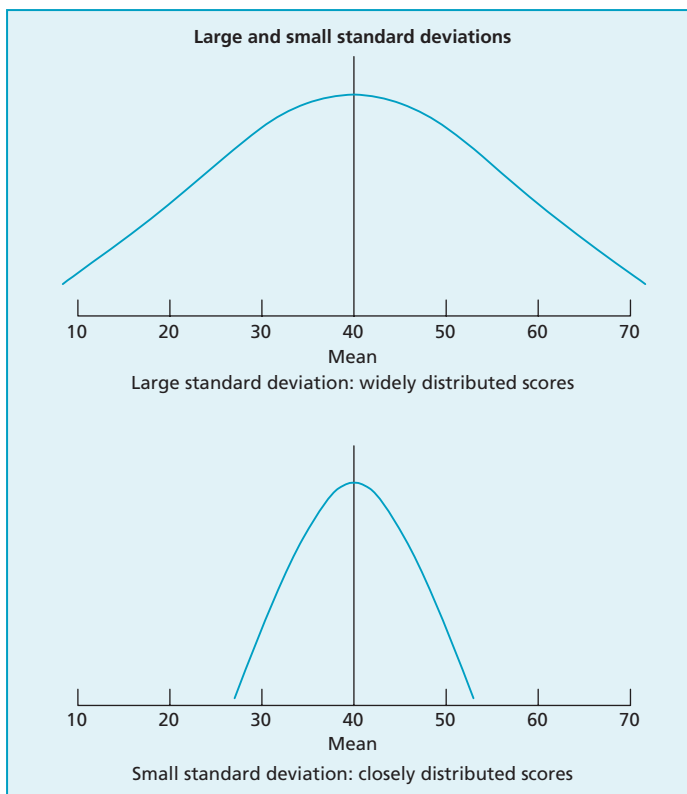
$$\text{Standard deviation} = \sqrt{11.33} = 3.37$$

The first step is to work out the mean of the sample. This is given by the total of all of the participants' scores ($\sum X = 130$; the symbol Σ means the sum of) divided by the number of participants ($N = 13$). Thus, the mean is 10.

KEY TERMS

Variation ratio: a measure of dispersion based on the proportion of the scores that are not at the modal value.

Standard deviation: a measure of dispersal that is of special relevance to the normal distribution; it is the square root of the variance. It takes account of every score, and is a sensitive dispersion measure.



The second step is to subtract the mean in turn from each score ($X - M$). The calculations are shown in the fourth column. The third step is to square each of the scores in the fourth column $(X - M)^2$. The fourth step is to work out the total of all the squared scores, $\Sigma(X - M)^2$. This comes to 136. The fifth step is to divide the result of the fourth step by one less than the number of participants, $N - 1 = 12$. This gives us 136 divided by 12, which equals 11.33. This is known as the **variance**, which is in squared units. Finally, we use a calculator to take the square root of the variance. This produces a figure of 3.37; this is the standard deviation.

The method for calculating the standard deviation that has just been described is used when we want to estimate the standard deviation of the population. If we want merely to describe the spread of scores in our sample, then the fifth step involves dividing the result of the fourth step by N .

What is the meaning of this figure for the standard deviation? We expect about two-thirds of the scores in a sample to lie within one standard deviation of the mean. In our example, the mean is 10.0, one standard deviation above the mean is 13.366 and one standard deviation below the mean is 6.634. In fact, 61.5% of the scores lie between those two limits, which is only slightly below the expected percentage.

The standard deviation has special relevance in relation to the so-called normal distribution. As was mentioned earlier, the normal distribution is a bell-shaped curve in which there are as many scores above the mean as below it. Intelligence (or IQ) scores in the general population provide an example of a normal distribution. Other characteristics such as height and weight also form roughly a normal distribution. Most of the scores in a normal distribution cluster fairly close to the mean, and there are fewer and fewer scores as you move away from the mean in either direction. In a normal distribution, 68.26% of the scores fall within one standard deviation of the mean, 95.44% fall within two standard deviations, and 99.73% fall within three standard deviations.

The standard deviation takes account of all of the scores and provides a sensitive measure of dispersion. As we have seen, it also has the advantage that it describes the spread of scores in a normal distribution with great precision. The most obvious disadvantage of the standard deviation is that it is much harder to work out than the other measures of dispersion.

DATA PRESENTATION

Information about the scores in a sample can be presented in several ways. If it is presented in a graph or chart, this may make it easier for people to understand what has been found, compared to simply presenting information about the central tendency and dispersion. We will shortly consider some examples. The key point to remember is that all graphs and charts should be clearly labelled and presented so that the reader can rapidly make sense of the information contained in them.

KEY TERM

Variance: a measure of dispersion that is the square of the standard deviation.

25 athletes running 400 metres

Raw data

Athlete	1	2	3	4	5	6	7	8	9
Speed	71	77	84	49	63	62	56	67	52
Athlete	10	11	12	13	14	15	16	17	18
Speed	61	63	59	48	61	65	68	54	61
Athlete	19	20	21	22	23	24	25		
Speed	58	66	55	57	58	56	53		

Table of frequencies (number of athletes obtaining each speed)

Speed	48	49	52	53	54	55	56	57	58	59	61	62	63	65	66	67	68	71	77	84
Athlete no.	13	4	9	25	17	21	7	22	19	12	10	6	5	15	20	8	16	1	2	3
							24	23	23	14	18	11								
Number	1	1	1	1	1	1	2	1	2	1	3	1	2	1	1	1	1	1	1	1

Suppose that we ask 25 male athletes to run 400 metres as rapidly as possible, and record their times (in seconds). Having worked out a table of frequencies (see the boxed example above), there are several ways to present these data.

Frequency polygon

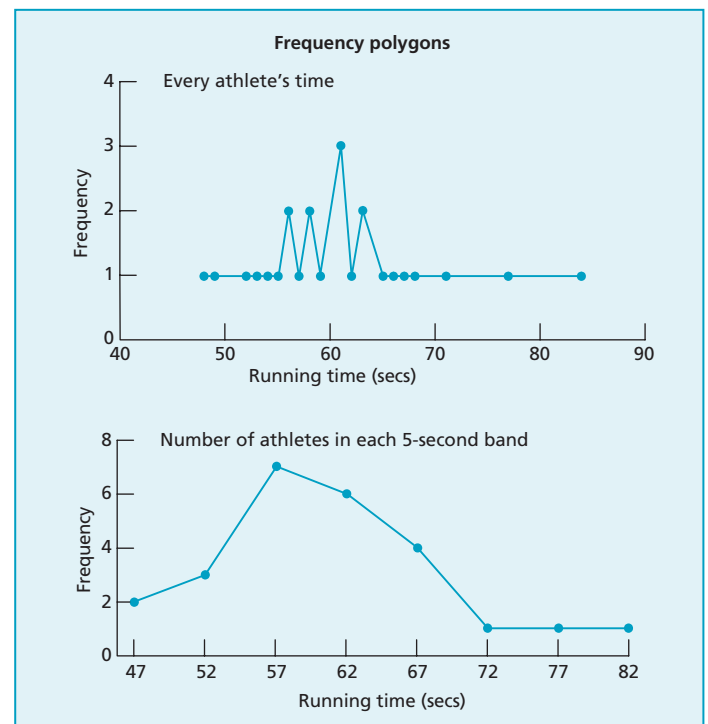
One way of summarising these data is in the form of a **frequency polygon**. This is a simple form of chart in which the scores from low to high are indicated on the x or horizontal axis and the frequencies of the various scores (in terms of the numbers of individuals obtaining each score) are indicated on the y or vertical axis. The points on a frequency polygon should only be joined up when the scores can be ordered from low to high. In order for a frequency polygon to be most useful, it should be constructed so that most of the frequencies are neither very high nor very low. The frequencies will be very high if the width of each class interval (the categories used to summarise frequencies) on the x axis is too broad (e.g. covering 20 seconds), and the frequencies will be very low if each class interval is too narrow (e.g. covering only 1 or 2 seconds).

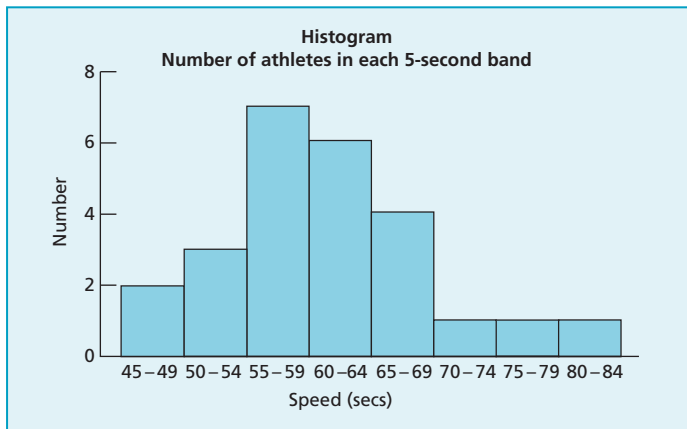
Each point in a frequency polygon should be placed in the middle of its class interval. There is a technical point that needs to be made here (Coolican, 1994). Suppose that we include all times between 53 and 57 seconds in the same class interval. As we have only measured running times to the nearest second, this class interval will cover actual times between 52.5 and 57.5 seconds. In this case, the mid-point of the class interval (55 seconds) is the same whether we take account of the actual measurement interval (52.5–57.5 seconds) or adopt the simpler approach of focusing on the lowest and highest recorded times in the class interval (53–57 seconds, respectively). When the two differ, it is important to use the actual measurement interval.

How should we interpret the findings shown in the frequency polygon? It is clear that most of the participants were able to run 400 metres in between about 53 and 67 seconds. Only a few of the athletes were able to better a time of 53 seconds, and there was a small number who took longer than 67 seconds.

KEY TERM

Frequency polygon: a graph showing the frequencies with which different scores are obtained by the participants in a study.





Histogram

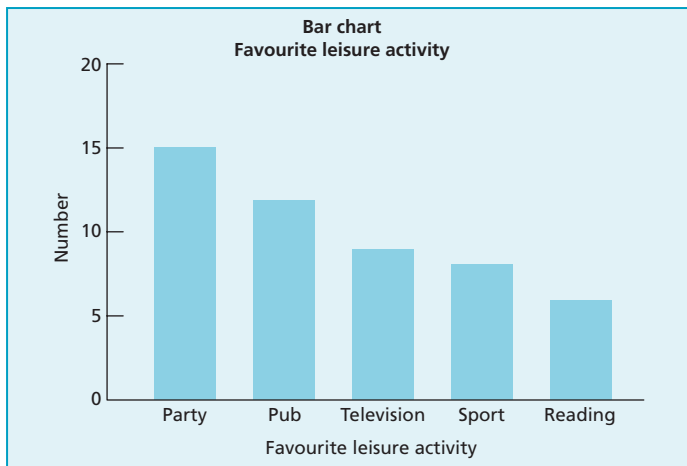
A similar way of describing these data is by means of a **histogram**. In a histogram, the scores are indicated on the horizontal axis and the frequencies are shown on the vertical axis. In contrast to a frequency polygon, however, the frequencies are indicated by rectangular columns. These columns are all the same width but vary in height in accordance with the corresponding frequencies. As with frequency polygons, it is important to make sure that the class intervals are not too broad or too narrow. All class intervals are represented, even if there are no scores in some of them. Class intervals are indicated by their mid-point at the centre of the columns.

Histograms are clearly rather similar to frequency polygons. However, frequency polygons are sometimes preferable when you want to compare two different frequency distributions.

The information contained in a histogram is interpreted in the same way as the information in a frequency polygon. In the present example, the histogram indicates that most of the athletes ran 400 metres fairly quickly. Only a few had extreme times

Bar chart

Frequency polygons and histograms are suitable when the scores obtained by the participants can be ordered from low to high. In more technical terms, the data should be either interval or ratio (see next section). However, there are many studies in which the scores are in the form of categories rather than ordered scores; in other words, the data are nominal. For example, 50 people might be asked to indicate their favourite leisure activity. Suppose that 15 said going to a party, 12 said going to the pub, 9 said watching television, 8 said playing sport, and 6 said reading a good book.



These data can be displayed in the form of a **bar chart**. In a bar chart, the categories are shown along the horizontal axis, and the frequencies are indicated on the vertical axis. In contrast to the data contained in histograms, the categories in bar charts cannot be ordered numerically in a meaningful way. However, they can be arranged in ascending (or descending) order of popularity. Another difference from histograms is that the rectangles in a bar chart do not usually touch each other.

The scale on the vertical axis of a bar chart normally starts at zero. However, it is sometimes convenient for presentational purposes to have it start at some higher value. If that is done, then it should be made clear in the bar chart that the lower part of the vertical scale is missing. The columns in a bar chart often represent frequencies. However, they can also represent means or percentages for different groups (Coolican, 1994).

How should we interpret the information in a bar chart? In the present example, a bar chart makes it easy to compare the popularity of different leisure activities. We can see at a glance that going to a party was the most popular leisure activity, whereas reading a good book was the least popular.

KEY TERM

Histogram: a graph in which the frequencies with which different scores are obtained by the participants in a study are shown by rectangles of different heights.

Bar chart: a graph showing the frequencies with which the participants in a study fall into different categories.

STATISTICAL TESTS

The various ways in which the data from a study can be presented are all useful in that they give us convenient and easily understood summaries of what we have found. However, to have a clearer idea of what our findings mean, it is generally necessary to

carry out one or more statistical tests. The first step in choosing an appropriate statistical test is to decide whether your data were obtained from an experiment in which some aspect of the situation (the independent variable) was manipulated in order to observe its effects on the dependent variables (i.e. the scores). If so, you need a test of difference (see pp. 17–22 of this chapter). On the other hand, if you simply have two observations from each of your participants in a non-experimental design, then you need a test of association or correlation (see pp. 23–25 of this chapter).

In using a statistical test, you need to take account of the experimental hypothesis. If you predicted the direction of any effects (e.g. loud noise will disrupt learning and memory), then you have a directional hypothesis, which should be evaluated by a one-tailed test. If you did not predict the direction of any effects (e.g. loud noise will affect learning and memory), then you have a non-directional hypothesis, which should be evaluated by a two-tailed test (see the Research methods: Design of investigations chapter).

Another factor to consider when deciding which statistical test to use is the type of data you have obtained. There are four types of data of increasing levels of precision:

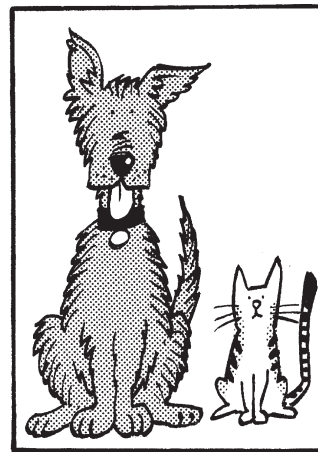
- **Nominal:** the data consist of the numbers of participants falling into various categories (e.g. fat, thin; men, women).
- **Ordinal:** the data can be ordered from lowest to highest (e.g. the finishing positions of athletes in a race).
- **Interval:** the data differ from ordinal data, because the units of measurement are fixed throughout the range; for example, there is the same “distance” between a height of 1.82 metres and 1.70 metres as between a height of 1.70 metres and one of 1.58 metres.
- **Ratio:** the data have the same characteristics as interval data, with the exception that they have a meaningful zero point; for example, time measurements provide ratio data because the notion of zero time is meaningful, and 10 seconds is twice as long as 5 seconds. The similarities between interval and ratio data are so great that they are sometimes combined and referred to as interval/ratio data.

Statistical tests can be divided into **parametric tests** and **non-parametric tests**. Parametric tests should only be used when the data obtained from a study satisfy various requirements. More specifically, there should be interval or ratio data, the data should be normally distributed, and the variances in the two conditions should be reasonably similar. In contrast, non-parametric tests can nearly always continue to be used, even when the requirements of parametric tests are satisfied. In this chapter, we will confine ourselves to a discussion of some of the most useful non-parametric tests.

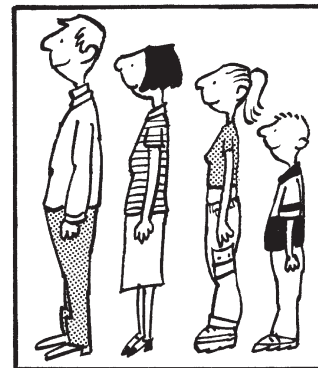
Statistical significance

So far we have discussed some of the issues that influence the choice of statistical test. What happens after we have chosen a statistical test, and analysed our data, and want to interpret our findings? We use the results of the test to choose between the following:

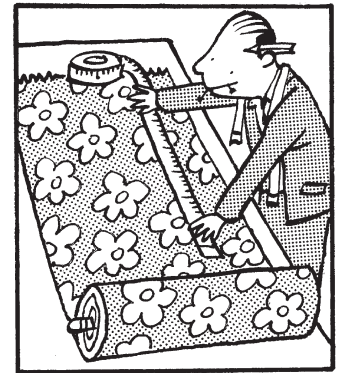
- Experimental hypothesis (e.g. loud noise disrupts learning).
- Null hypothesis, which asserts that there is no difference between conditions (e.g. loud noise has no effect on learning).



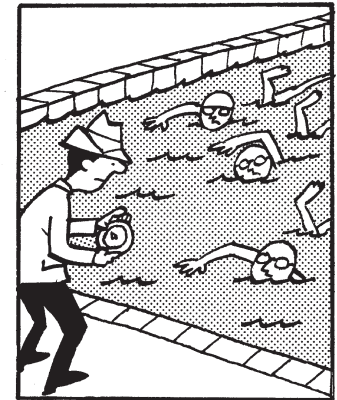
Nominal



Ordinal



Interval



Ratio

KEY TERMS

Nominal data: data consisting of the numbers of participants falling into qualitatively different categories.

Ordinal data: data that can be ordered from smallest to largest.

Interval data: data in which the units of measurement have an invariant or unchanging value.

Ratio data: as interval data, but with a meaningful zero point.

Parametric tests: statistical tests that require interval or ratio data, normally distributed data, and similar variances in both conditions.

Non-parametric tests: statistical tests that do not involve the requirements of parametric tests.

■ Research activity: Devising hypotheses

Devise suitable null and experimental hypotheses for the following:

- An investigator considers the effect of noise on students' ability to concentrate and complete a word-grid. One group only is subjected to the noise in the form of a distractor, i.e. a television programme.
- An investigator explores the view that there might be a link between the amount of television children watch and their behaviour at school.

From percentage to decimal

10% = 0.10
 5% = 0.05
 1% = 0.01
 2.5% = ?

To go from decimal to percentage, multiply by 100: move the decimal point two places to the right.

To go from percentage to decimal, divide by 100: move the decimal point two places to the left.

If the statistical test indicates that there is only a small probability of the difference between conditions (e.g. loud noise vs. no noise) having occurred if the null hypothesis were true, then we reject the null hypothesis in favour of the experimental hypothesis.

Why do we focus initially on the null hypothesis rather than the experimental hypothesis? The reason is that the experimental hypothesis is rather imprecise. It may state that loud noise will disrupt learning, but it does not indicate the *extent* of the disruption. This imprecision makes it hard to evaluate an experimental hypothesis directly. In

contrast, a null hypothesis such as loud noise has no effect on learning is precise, and this precision allows us to use statistical tests to decide the probability that it is correct.

Psychologists generally use the 5% (0.05) level of **statistical significance**. What this means is that the null hypothesis is rejected (and the experimental hypothesis is accepted) if the probability that the results were due to chance alone is 5% or less. This is often expressed as $p = 0.05$, where p = the probability of the result if the null hypothesis is true. If the statistical test indicates that the findings do not reach the 5% (or $p = 0.05$) level of statistical significance, then we retain the null hypothesis, and reject the experimental hypothesis. The key decision is whether or not to reject the null hypothesis and that is why the 0.05 level of statistical significance is so important. However, our data sometimes indicate that the null hypothesis can be rejected with greater confidence, say, at the 1% (0.01) level. If the null hypothesis can be rejected at the 1% level, it is customary to state that the findings are highly significant. In general terms, you should state the precise level of statistical significance of your findings, whether it is the 5% level, the 1% level, or whatever.

These procedures may seem easy. In fact, there are two errors that may occur when reaching a conclusion on the basis of the results of a statistical test:

- **Type I error:** we may reject the null hypothesis in favour of the experimental hypothesis even though the findings are actually due to chance; the probability of this happening is given by the level of statistical significance that is selected.
- **Type II error:** we may retain the null hypothesis even though the experimental hypothesis is actually correct.

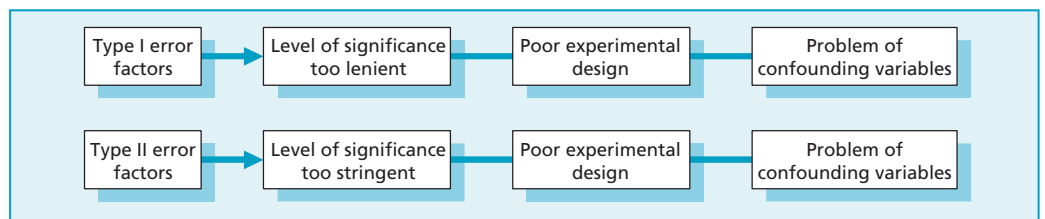
It would be possible to reduce the likelihood of a Type I error by using a more stringent level of significance. For example, if we used the 1% ($p = 0.01$) level of significance, this would greatly reduce the probability of a Type I error. However, use of a more stringent level of significance increases the probability of a Type II error. We could reduce the probability of a Type II error by using a less stringent level of significance, such as the 10% ($p = 0.10$) level. However, this would increase the probability of a Type I error. These considerations help to make it clear why most psychologists favour the 5% (or $p = 0.05$) level of significance: it allows the probabilities of both Type I and Type II errors to remain reasonably low.

KEY TERMS

Statistical significance: the level at which the decision is made to reject the null hypothesis in favour of the experimental hypothesis.

Type I error: mistakenly rejecting the null hypothesis in favour of the experimental hypothesis when the results are actually due to chance.

Type II error: mistakenly retaining the null hypothesis when the experimental hypothesis is actually correct.



Psychologists generally use the 5% level of significance. However, they would use the 1% or even the 0.1% level of significance if it were very important to avoid making a Type I error. For example, clinical psychologists might require very strong evidence that a new form of therapy was more effective than existing forms of therapy before starting to use it on a regular basis. The 1% or 0.1% ($p = 0.001$) level of statistical significance is also used when the experimental hypothesis seems improbable. For example, very few people would accept that telepathy had been proved to exist on the basis of a single study in which the results were only just significant at the 5% level!

Tests of difference

In this section, we will consider those statistical tests that are applicable when we are interested in deciding whether the differences between two conditions or groups are significant. As discussed in the Research methods: Design of investigations chapter, there are three kinds of design that can be used when we want to compare two conditions. First, there is the independent design, in which each participant is allocated at random to one and only one condition. Second, there is the repeated measures design, in which the same participants are used in both conditions. Third, there is the matched participants design, in which the participants in the two conditions are matched in terms of some variable or variables that might be relevant (e.g. intelligence; age).

When deciding which statistical test to use, it is very important to take account of the particular kind of experimental design that was used. If the independent design has been used, then the Mann-Whitney U test is likely to be an appropriate test to use. If the repeated measures or matched participants design has been used, then the sign test or the Wilcoxon matched pairs signed ranks test is likely to be appropriate. Each of these tests is discussed in turn next.

Mann-Whitney U test

The Mann-Whitney U test can be used when an independent design has been used, and the data are either ordinal or interval. The worked example in the box shows how this test is calculated.

Mann-Whitney U test: A worked example

Experimental hypothesis: extensive training improves performance

Null hypothesis: training has no effect on performance

Participant	Condition A	Rank	Participant	Condition B	Rank
1	4	2	1	21	15
2	10	9	2	26	18
3	12	11	3	20	14
4	28	20	4	22	16
5	7	5	5	32	22
6	13	13	6	5	3
7	12	11	7	12	11
8	2	1	8	6	4
9	9	7.5	9	8	6
10	27	19	10	24	17
			11	29	21
			12	9	7.5

Smaller sample = condition A

Sum of ranks in smaller sample (T) = 98.5

Number of participants in smaller sample (N_A) = 10

Number of participants in larger sample (N_B) = 12

Formula: $U = N_A N_B + \left(\frac{N_A(N_A + 1)}{2} \right) - T$

Example: $U = (10 \times 12) + \left(\frac{10(10 + 1)}{2} \right) - 98.5 = 76.5$

Formula for calculating U' : $U' = N_A N_B - U$

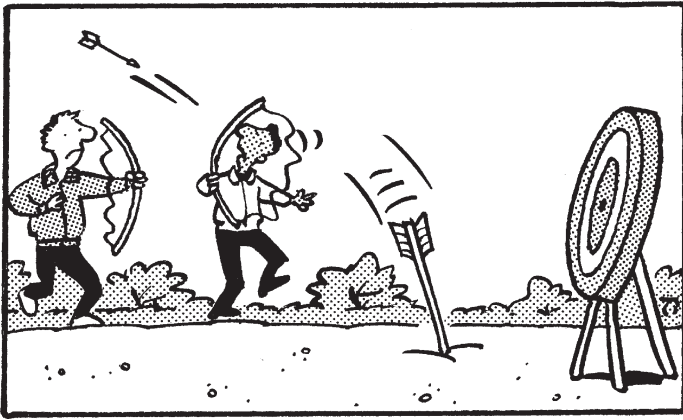
Example: $U' = (10 \times 12) - 76.5 = 43.5$

Comparing U and U' , U' is the smaller value. The calculated value of U' (43.5) is checked against the tabled value for a one-tailed test at 5%.

Table values

	$N_A = 10$
$N_B = 12$	34

Conclusion: as 43.5 is greater than 34, the null hypothesis should be retained—i.e. training has no effect on performance in this task.



Suppose that we have two conditions. In both conditions, the participants have to fire arrows at a board, and the score obtained is recorded. There are 10 participants in Condition A, in which no training is provided before their performance is assessed. There are 12 participants in Condition B, and they receive extensive training before their performance is assessed. The experimental hypothesis was that extensive training would improve performance; in other words, the scores in Condition B should be significantly higher than those in Condition A.

The first step is to rank all of the scores from both groups together, with a rank of 1 being given to the smallest score, a rank of 2 to the second smallest score, and so on. If there are tied scores, then the mean of the ranks

involved is given to each of the tied participants. For example, two participants were tied for the 7th and 8th ranks, and so they both received a rank of 7.5.

The second step is to work out the sum of the ranks in the smaller sample, which is Condition A in our example. This value is known as T , and it is 98.5 in the example.

The third step is to calculate U from the formula

$$U = N_A N_B + \left(\frac{N_A(N_A + 1)}{2} \right) - T,$$

in which N_A is the number of participants in the smaller sample and N_B is the number in the larger sample.

The fourth step is to calculate U' from the formula $U' = N_A N_B - U$.

The fifth step is to compare U and U' , selecting whichever is the smaller value provided that the results are in the correct direction. The smaller value (i.e. 43.5) is then looked up in Appendix 1. The observed value must be equal to, or smaller than, the tabled value in order to be significant. In this case, we have a one-tailed test, because the experimental hypothesis stated that extensive training would improve performance and the statistical significance is the standard 5% (0.05). With 10 participants in our first condition and 12 in our second condition, the tabled value for significance is 34 (value obtained from the table at the bottom of page 5 of the Appendices). As our value of 43.5 is greater than 34, the conclusion is that we retain the null hypothesis. It should be noted that the presence of ties reduces the accuracy of the tables, but the effect is small unless there are several ties.

Sign test

The sign test can be used when a repeated measures or matched participants design has been used, and the data are ordinal. If the data are interval or ratio, then it would be more appropriate to use the Wilcoxon matched pairs signed ranks test. The worked example in the box illustrates the way in which the sign test is calculated.

The sign test is ideal to use if the data are ordinal as it analyses at a very basic level, e.g. in a race it can tell you that "John beat Peter". It can also be used with interval or ratio data, but as it only gives a crude analysis, this data would be better applied to the Wilcoxon test, which can give a more sophisticated analysis, e.g. "John beat Peter by 2 seconds".

Suppose that there were 12 participants in an experiment. In Condition A these participants were presented with 20 words to learn in a situation with no noise; learning was followed five minutes later by a test of free recall in which they wrote down as many words as they could

remember in any order. Condition B involved presenting 20 different words to learn in a situation of loud noise, again followed by a test of free recall. The experimenter predicted that free recall would be higher in the no-noise condition. Thus, there was a directional hypothesis.

In order to calculate the sign test it is necessary first of all to draw up a table like the one in the example, in which each participant's scores in Condition A and in Condition B are recorded. Each participant whose score in Condition A is greater than his or her score

Sign test: A worked example

Experimental hypothesis: free recall is better when learning takes place in the absence of noise than in its presence

Null hypothesis: free recall is not affected by whether or not noise is present during learning

Participant	Condition A (no noise)	Condition B (loud noise)	Sign
1	12	8	+
2	10	10	0
3	7	8	-
4	12	11	+
5	8	3	+
6	10	10	0
7	13	7	+
8	8	9	-
9	14	10	+
10	11	9	+
11	15	12	+
12	11	10	+

Number of + signs = 8

Number of - signs = 2

Number of 0 signs = 2

Number of participants with differing scores (N) = $8 + 2 = 10$

Number of participants with less-frequent sign (S) = 2

Question: Is the value of S in this example the same as or lower than the tabled value for S ?

Table values

	5%
$N = 10$	$S = 1$

Conclusion: in this experiment the value of S is higher than the tabled value, when $N = 10$. The null hypothesis (that noise has no effect on learning and memory) cannot be rejected.

in Condition B is given a plus sign (+) in the sign column, and each participant whose score in Condition B is greater than his or her score in Condition A is given a minus sign (-) in the sign column. Each participant whose scores in both conditions are the same receives a 0 sign in the sign column, and are ignored in the subsequent calculations—they do not contribute to N (the number of paired scores), as they provide no evidence about effect direction.

In the example, there are eight plus signs, two minus signs, and two participants had the same scores in both conditions. If we ignore the two participants with the same scores in both conditions, this gives us $N = 10$. Now all we need to do is to work out the number of these 10 participants having the less frequently occurring sign; this value is known as S . In terms of our example, $S = 2$. We can refer to the relevant table (Appendix 2) with $N = 10$ and $S = 2$ and the statistical significance is the standard 5%. The obtained value for S must be the same as or lower than the value for S given in the table. The tabled value for a one-tailed test is 1. Thus, our obtained S value of 2 is not significant at the 5% level on a one-tailed test. We therefore conclude that we cannot reject the null hypothesis that noise has no effect on learning and memory.

Wilcoxon matched pairs signed ranks test

The Wilcoxon matched pairs signed ranks test can be used when a repeated measures or matched participants design has been used, and the data are at least ordinal. This test or the sign test can be used if the data are ordinal, interval, or ratio. However, the Wilcoxon matched pairs signed ranks test uses more of the information obtained from a study, and so is usually a more sensitive and useful test than the sign test.

The worked example uses the data from the sign test. The first step is to place all the data in a table in which each participant's two scores are in the same row. The second step

Use the sign test if your data are ordinal, and the Wilcoxon test if your data are interval or ratio to get the best results from your analysis.

Wilcoxon matched pairs signed ranks test: A worked example

Experimental hypothesis: free recall is better when learning takes place in the absence of noise than in its presence

Null hypothesis: free recall is not affected by whether or not noise is present during learning

Participant	Condition A (no noise)	Condition B (loud noise)	Difference (d) (A – B)	Rank
1	12	8	4	7.5
2	10	10	0	–
3	7	8	–1	2.5
4	12	11	1	2.5
5	8	3	5	9
6	10	10	0	–
7	13	7	6	10
8	8	9	–1	2.5
9	14	10	4	7.5
10	11	9	2	5
11	15	12	3	6
12	11	10	1	2.5

Sum of positive ranks (7.5 + 2.5 + 9 + 10 + 7.5 + 5 + 6 + 2.5) = 50

Sum of negative ranks (2.5 + 2.5) = 5

Smaller value (5) = T

Number of participants who scored differently in condition A and B (N) = 10

Question: For the results to be significant, the value of T must be the same as, or less than, the tabled value.

Table values

	5%	1%
N = 10	11	5

Conclusion: in this experiment T is less than the tabled value at the 5% level and the same as the tabled value at the 1% level of significance, so the null hypothesis is rejected in favour of the experimental hypothesis.

is to subtract the Condition B score from the Condition A score for each participant to give the difference (d). The third step is to omit all the participants whose two scores are the same, i.e. $d = 0$. The fourth step is to rank all the difference scores obtained in the second step from 1 for the smallest difference, 2 for the second smallest difference, and so on. For this purpose, ignore the + and – signs, thus taking the absolute size of the difference. The fifth step is to add up the sum of the positive ranks (50 in the example) and separately to add up the sum of the negative ranks (5 in the example). The smaller of these values is T, which in this case is 5. The sixth step is to work out the number of participants whose two scores are not the same, i.e. $d \neq 0$. In the example, $N = 10$.

The obtained value of T must be the same as, or less than, the tabled value (see Appendix 3) in order for the results to be significant. The tabled value for a one-tailed test and $N = 10$ is 11 at the 5% level of statistical significance, and it is 5 at the 1% level. Thus, the findings are significant at the 1% level on a one-tailed test. The null hypothesis is rejected in favour of the experimental hypothesis that free recall is better when learning takes place in the absence of noise than in its presence ($p = 0.01$). The presence of ties means that the tables are not completely accurate, but this does not matter provided that there are only a few ties.

You may be wondering how it is possible for the same data to produce a significant finding on a Wilcoxon matched pairs signed ranks test but not on a sign test. Does this indicate that statistics are useless? Not at all. The sign test is insensitive (or lacking in power) because it takes no account of the *size* of each individual's difference in free recall in the two conditions. It is because this information is made use of in the Wilcoxon matched pairs signed ranks test that a significant result was obtained using that test. Thus, the Wilcoxon matched pairs signed ranks test has more power than the sign test to detect differences between two conditions.

Correlational studies

In the case of correlational studies, the data are in the form of two measures of behaviour from each member of a single group of participants. What is often done is to present the data in the form of a **scattergraph** (also known as a scattergram). It is given this name, because it shows the ways in which the scores of individuals are scattered.

Scattergraphs

Suppose that we have carried out a study on the relationship between the amount of television violence seen and the amount of aggressive behaviour displayed. We could have a scale of the amount of television violence seen on the horizontal axis, and a scale of the amount of aggressive behaviour on the vertical axis. We could then put a dot for each participant indicating where he or she falls on these two dimensions. For example, suppose that one individual watched 17 hours of television and obtained a score of 8 for aggressive behaviour. We would put a cross at the point where the invisible vertical line from the 17 meets the invisible horizontal line from the 8.

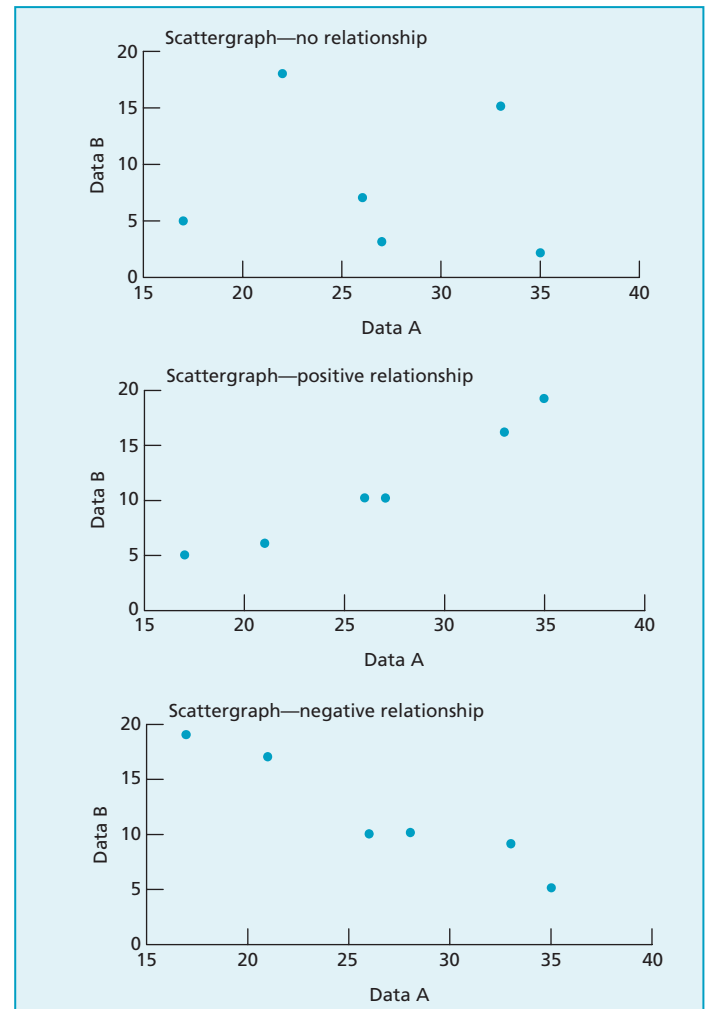
How do we interpret the information contained in a scattergraph? If there is a positive relationship between watching violence and aggression, then the dots should tend to form a pattern going from the bottom left of the scattergraph to the top right. If there is no relationship between the two variables, then the dots should be distributed in a fairly random way within the scattergraph. If there is a negative relationship between the two variables, then the dots will form a pattern going from the top left to the bottom right. In the present example, this would mean that watching a lot of television violence was associated with a *low* level of aggression.

As we will see shortly, the strength of a correlation between two variables can be assessed statistically by Spearman's rho. What, then, is the value of a scattergraph? Spearman's rho is limited in that it sometimes indicates that there is no relationship between two variables even when there is. For example, Spearman's rho would not reveal the existence of a strong curvilinear relationship between two variables, but this would be immediately obvious in a scattergraph.

Spearman's rho

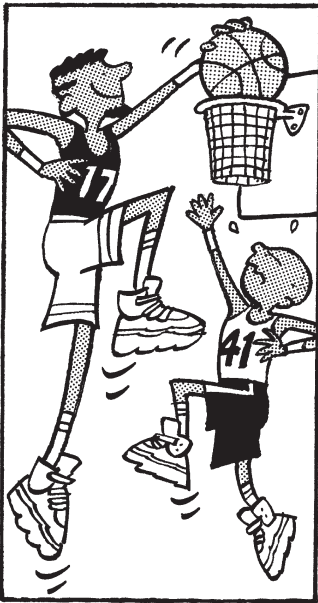
Suppose that we have scores on two variables from each of our participants, and we want to see whether there is an association or correlation between the two sets of scores. This can be done by using a test known as Spearman's rho, provided that the data are at least ordinal. Spearman's rho or r_s indicates the strength of the association. If r_s is $+1.0$, then there is a perfect positive correlation between the two variables. If r_s is -1.0 , then there is a perfect negative correlation between the two variables. If r_s is 0.0 , then there is generally no relationship between the two variables. The working of this test is shown in the worked example.

An experimenter collects information about the amount of television violence seen in the past month and about the amount of aggressive behaviour exhibited in the past month from 12 participants. She predicts that there will be a positive association between these two variables, i.e. those participants who have seen the most television violence (variable A) will tend to be the most aggressive (variable B). In other words, there is a directional hypothesis.

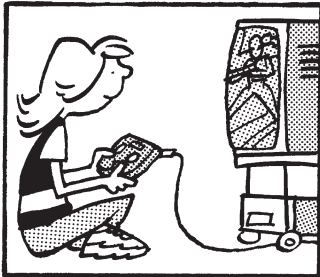


KEY TERM

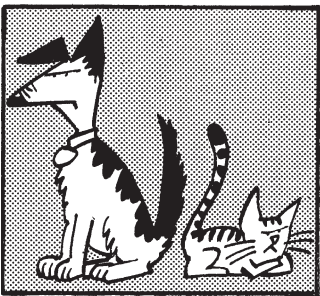
Scattergraph: two-dimensional representation of all the participants' scores in a correlational study; also known as scattergram.



*A positive correlation:
The taller the player,
the higher the score.*



*A negative correlation:
The more time spent
playing computer games,
the less time spent
studying.*



*No correlation:
Where there is no
relationship, variables
are uncorrelated.*

The first step is to draw up a table in which each participant's scores for the two variables are placed in the same row.

The second step is to rank all the scores for variable A. A rank of 1 is assigned to the smallest score, a rank of 2 to the second smallest score, and so on up to 12. What do we do if there are tied scores? In the example, participants 9 and 12 had the same score for variable A. The ranks that they are competing for are ranks 5 and 6. What is done is to take the average or mean of the ranks at issue: $(5 + 6)/2 = 5.5$.

The third step is to rank all the scores for variable B, with a rank of 1 being assigned to the smallest score. Participants 6, 7, 9, and 11 are all tied, with the ranks at issue being ranks 4, 5, 6, and 7. The mean rank at issue is $(4 + 5 + 6 + 7)/4 = 5.5$.

The fourth step is to calculate the difference between the two ranks obtained by each individual, with the rank for variable B being subtracted from the rank for variable A. This produces 12 difference (d) scores.

The fifth step is to square all of the d scores obtained in the fourth step. This produces 12 squared difference (d^2) scores.

The sixth step is to add up all of the d^2 scores in order to obtain the sum of the squared difference scores. This is known as $\sum d^2$, and comes to 30 in the example.

The seventh step is to work out the number of participants. In the example, the number of participants (N) is 12.

A worked example of a test for correlation between two variables using Spearman's rho

Experimental hypothesis: there is a positive association between amount of television violence watched and aggressive behaviour

Null hypothesis: there is no association between amount of television violence watched and aggressive behaviour

Participants	TV violence seen (hours)	Aggressive behaviour (out of 10)	Rank A	Rank B	Difference d	d^2
1	17	8	7.5	9	-1.50	2.25
2	6	3	2	2	0.00	0.00
3	23	9	10	10.5	-0.50	0.25
4	17	7	7.5	8	-0.50	0.25
5	2	2	1	1	0.00	0.00
6	20	6	9	5.5	+3.50	12.25
7	12	6	4	5.5	0.00	2.25
8	31	10	12	12	0.00	0.00
9	14	6	5.5	5.5	+0.50	0.00
10	26	9	10.5	10.5	+0.50	0.25
11	9	6	5.5	5.5	-2.50	6.52
12	14	4	3	3	+2.50	6.25

Sum of squared difference scores ($\sum d^2$) = 30

Number of participants (N) = 12

Formula: $\rho = 1 - \frac{(\sum d^2 \times 6)}{N(N^2 - 1)}$

Example: $1 - \frac{(30 \times 6)}{12(143)} = 1 - 0.105 = +0.895$

Is the value of rho (+0.895) as great as, or greater than the tabled value?

Table values

	0.05 level	0.01 level	0.005 level
N = 12	+0.503	+0.671	+0.727

Conclusion: null hypothesis rejected in favour of experimental hypothesis, i.e. there is a positive correlation between the amount of television violence watched and aggressive behaviour ($p = 0.005$).

The eighth step is to calculate rho from the following formula:

$$\text{rho} = 1 - \frac{(\sum d^2 \times 6)}{N(N^2 - 1)}$$

In the example, this becomes $1 - \frac{(30 \times 6)}{12(143)} = 1 - 0.105 = +0.895$

The ninth and final step is to work out the significance of the value of rho by referring the result to the table (see Appendix 4). The obtained value must be as great as, or greater than, the tabled value. The tabled value for a one-tailed test with $N = 12$ is +0.503 at the 0.05 level, it is +0.671 at the 0.01 level, and it is +0.727 at the 0.005 level. Thus, it can be concluded that the null hypothesis should be rejected in favour of the experimental hypothesis that there is a positive correlation between the amount of television violence watched and aggressive behaviour ($p = 0.005$).

An important point about Spearman's rho is that the statistical significance of the obtained value of rho depends very heavily on the number of participants. For example, the tabled value for significance at the 0.05 level on a one-tailed test is +0.564 if there are 10 participants. However, it is only +0.306 if there are 30 participants. In practical terms, this means that it is very hard to obtain a significant correlation with Spearman's rho if the number of participants is low.

Test of association

The **chi-squared test** is a test of association. It is used when we have nominal data in the form of frequencies, and when each and every observation is independent of all the other observations. For example, suppose that we are interested in the association between eating patterns and cholesterol level. We could divide people into those having a healthy diet with relatively little fat and those having an unhealthy diet. We could also divide them into those having a fairly high level of cholesterol and those having a low level of cholesterol. In essence, the chi-squared test tells us whether membership of a given category on one dimension (e.g. unhealthy diet) is associated with membership of a given category on the other dimension (e.g. high cholesterol level).

In the worked example, we will assume that we have data from 186 individuals with an unhealthy diet, and from 128 individuals with a healthy diet. Of those with an unhealthy diet, 116 have a high cholesterol level and 70 have a low cholesterol level. Of those with a healthy diet, 41 have a high cholesterol level and 87 have a low cholesterol level. Our experimental hypothesis is that there is an association between healthiness of diet and low cholesterol level.

The first step is to arrange the frequency data in a 2×2 "contingency table" as in the worked example, with the row and column totals included. The second step is to work out what the four frequencies would be if there were no association at all between diet and cholesterol levels. The expected frequency (by chance alone) in each case is given by the following formula:

$$\text{expected frequency} = \frac{\text{row total} \times \text{column total}}{\text{overall total}}$$

For example, the expected frequency for the number of participants having a healthy diet and high cholesterol is 157×128 divided by 314, which comes to 64. The four expected frequencies (those expected by chance alone) are also shown in the table.

The third step is to apply the following formula to the observed (O) and expected (E) frequencies in each of the four categories:

$$\frac{(|O - E| - 1/2)^2}{E}$$

The "6" in the equation is always present, and is a feature of the Spearman's rho formula.

According to APA convention, numbers that cannot be greater than 1, e.g. correlations and probabilities, should be presented without a zero before the decimal point. However, this convention has not been used throughout this text.

As vertical lines denote absolute values, " $|O - E|$ " is the difference between these values disregarding the sign. Whether it is $3 - 5$ or $5 - 3$, the difference is always a positive number, i.e. 2 in this case.

KEY TERM

Chi-squared test: a test of association that is used with nominal data in the form of frequencies.

In the formula, $|O - E|$ means that the difference between the observed and the expected frequency should be taken, and it should then have a + sign put in front of it regardless of the direction of the difference. The correction factor (i.e. $-1/2$) is only used when there are two rows and two columns.

The fourth step is to add together the four values obtained in the third step in order to provide the chi-squared statistic or χ^2 . This is $7.91 + 5.44 + 7.91 + 5.44 = 26.70$.

The fifth step is to calculate the number of “degrees of freedom” (df). This is given by (the number of rows $- 1$) \times (the number of columns $- 1$). For this we need to refer back to the contingency table. In the example, this is $1 \times 1 = 1$. Why is there 1 degree

Test of association: Chi-squared test, a worked example

Experimental hypothesis: there is an association between healthiness of diet and low cholesterol level

Null hypothesis: there is no association between healthiness of diet and low cholesterol level

Contingency table:

	Healthy diet	Unhealthy diet	Row total
High cholesterol	41	116	157
Low cholesterol	87	70	157
Column total	128	186	314

Expected frequency if there were no association:

Formula: $\frac{\text{row total} \times \text{column total}}{\text{overall total}} = \text{expected frequency}$

	Healthy diet	Unhealthy diet	Row total
High cholesterol	64	93	157
Low cholesterol	64	93	157
Column total	128	186	314

Calculating chi-squared statistic (χ^2):

Formula: $\chi^2 = \sum \frac{(|O - E| - 1/2)^2}{E} = 26.7$

Note: Correction factor ($-1/2$) is only used where there are two rows and two columns

Category	Observed	Expected	$ O - E $	$\frac{(O - E - 1/2)^2}{E}$
Healthy, high cholesterol	41	64	23	7.91
Unhealthy, high cholesterol	116	93	23	5.44
Healthy, low cholesterol	87	64	23	7.91
Unhealthy, low cholesterol	70	93	23	5.44
				26.70

Calculating degrees of freedom:

Formula: (no. of rows $- 1$) \times (no. of columns $- 1$) = degrees of freedom $(2 - 1) \times (2 - 1) = 1$

Compare chi-squared statistic with tabled values:

Table values

	0.025 level	0.005 level	0.0005 level
df = 1	3.84	6.64	10.83

Question: is the observed chi-square value of 26.70 and one degree of freedom the same as or greater than the tabled value?

Conclusion: the chi-square value is greater than the tabled value, so the null hypothesis can be rejected, and the experimental hypothesis, that there is an association between healthiness of diet and cholesterol level, accepted.

of freedom? Once we know the row and column totals, then only one of the four observed values is free to vary. Thus, for example, knowing that the row totals are 157 and 157, the column totals are 128 and 186, and the number of participants having a healthy diet and high cholesterol is 41, we can complete the entire table. In other words, the number of degrees of freedom corresponds to the number of values that are free to vary.

The sixth step is to compare the tabled values in Appendix 5 with $\chi^2 = 26.70$ and one degree of freedom. The observed value needs to be the same as, or greater than, the tabled value for a one-tailed test in order for the results to be significant.

The tabled value for a one-tailed test with $df = 1$ is 3.84 at the 0.025 level, 6.64 at the 0.005 level, and 10.83 at the 0.0005 level. Thus, we can reject the null hypothesis, and conclude that there is an association between healthiness of diet and cholesterol level ($p = 0.0005$).

It is easy to use the chi-squared test wrongly. According to Robson (1994), “There are probably more inappropriate and incorrect uses of the chi-square test than of all the other statistical tests put together.” In order to avoid using the chi-squared test wrongly, it is important to make use of the following rules:

- Ensure that every observation is independent of every other observation; in other words, each individual should be counted once and in only *one* category.
- Make sure that each observation is included in the appropriate category; it is not permitted to omit some of the observations (e.g. those from individuals with intermediate levels of cholesterol).
- The total sample should exceed 20; otherwise, the chi-squared test as described here is not applicable. More precisely, the minimum expected frequency should be at least 5 in every use.
- The significance level of a chi-squared test is assessed by consulting the one-tailed values in the Appendix table if a specific form of association has been predicted and that form was obtained. However, the two-tailed values should always be consulted if there are more than two categories on either dimension.
- Remember that showing that there is an association is not the same as showing that there is a causal effect; for example, the association between a healthy diet and low cholesterol does not demonstrate that a healthy diet *causes* low cholesterol.

ISSUES OF EXPERIMENTAL AND ECOLOGICAL VALIDITY

Assume that you have carried out a study, and then analysed it using a statistical test. The results were statistically significant, so you are able to reject the null hypothesis in favour of the experimental hypothesis. When deciding how to interpret your findings, you need to take account of issues relating to experimental and ecological validity. **Experimental validity** is based on the extent to which a given finding is genuine, and is due to the independent variable that was manipulated. In other words, it is essentially the same as internal validity, which is discussed in the Research methods: Design of investigations chapter. In contrast, **ecological validity** refers to the extent to which research findings can be generalised to a range of real-world settings. It is clearly desirable for a study to possess both of these forms of validity.

Experimental validity

How can we assess the experimental or internal validity of the findings from a study? The key point is made in the Research methods: Design of investigations chapter: we can only have confidence that the independent variable produced the observed effects on behaviour or the dependent variable provided that all of the principles of experimental design were followed. These principles include the standardisation of instructions and procedures; counterbalancing; randomisation; and the avoidance of confounding variables, experimenter effects, demand characteristics, and participant reactivity.

KEY TERMS

Experimental validity: the extent to which a finding is genuine, and due to the independent variable being manipulated.

Ecological validity: the extent to which the findings of laboratory studies generalise to other locations, times, and measures.

We can check these by asking various questions about a study, including the following:

- Were there any variables (other than the independent variable) that varied systematically between conditions?
- Did all the participants receive the same standardised instructions?
- Were the participants allocated at random to the conditions?
- Did the experimenter influence the performance of the participants by his or her expectations or biases?
- Were the participants influenced by any demand characteristics of the situation?
- If the participants knew they were being observed, did this influence their behaviour?

Probably the most convincing evidence that a study possesses good experimental validity is if its findings can be repeated or replicated in other studies. Why is that so? Suppose, for example, that we obtain significant findings in one study because we failed to allocate our participants at random to conditions. Anyone else carrying out the same study, but allocating the participants at random, would be very unlikely to repeat the findings of our study.

Ecological validity

As Coolican (1994) pointed out, the term ecological validity has been used in various ways. It is sometimes used to refer to the extent to which a given study was carried out in a naturalistic or real-world setting rather than an artificial one. However, as was mentioned earlier, it is probably more useful to regard ecological validity as referring to the extent to which a study generalises to various real-world settings. Bracht and Glass (1968) put forward a definition of ecological validity along those lines. According to them, the findings of ecologically valid studies generalise to other locations or places, to other times, and to other measures. Thus, the notion of ecological validity closely resembles that of external validity (see the Research methods: Design of investigations chapter), except that external validity also includes generalisation to other populations.

How do we know whether the findings of a study possess ecological validity? The only conclusive way of answering that question is by carrying out a series of studies in different locations, at different times, and using different measures. Following that approach is generally very costly in terms of time and effort.

It is often possible to obtain some idea of the ecological validity of a study by asking yourself whether there are important differences between the way in which a study has been conducted and what happens in the real world. For example, consider research on eyewitness testimony (see PIP, Chapter 9). The participants in most laboratory studies of eyewitness testimony have been asked to pay close attention to a series of slides or a video depicting some incident, after which they are asked various questions. The ecological validity of such studies is put in danger for a number of reasons. The participants have their attention directed to the incident, whereas eyewitnesses to a crime or other incident may fail to pay much attention to it. In addition, eyewitnesses are often very frightened and concerned about their own safety, whereas the participants in a laboratory study are not.

CASE STUDY: *Criticism of Intelligence Testing*

Gould's (1982) study included criticism of intelligence testing based on the methodological and theoretical problems experienced when these tests are used. Gould suggested that many IQ tests contain errors of validity. They have design flaws in relation to the wording used, which is often based on cultural definitions of meaning. Lack of access to the relevant cultural interpretations would disadvantage certain groups and individuals. For example, the Yerkes Tests of Intelligence were based on American culture and cultural knowledge, so that immigrants' performance was almost always

poorer than that of the native groups. Gould also emphasised the fact that the procedures used were flawed, especially during the testing of black participants.

Interpretation of findings from the use of Yerkes tests ignored the role of experience and education in IQ, and focused on the role of heredity. The research evidence was used to support racist social policy, which restricted work opportunities for ethnic groups within society and denied many the right to seek political refuge in America.

It may seem reasonable to argue that we could ensure ecological validity by taking research out of the laboratory and into the real world. However, powerful arguments against doing that with memory research were put forward by Banaji and Crowder (1989):

Imagine astronomy being conducted with only the naked eye, biology without tissue cultures ... or chemistry without test tubes! The everyday world is full of principles from these sciences in action, but do we really think their data bases should have been those of everyday applications? Of course not. Should the psychology of memory be any different? We think not.

In sum, investigators should consider the issue of ecological validity seriously when interpreting their findings. They should try to identify the main ways in which the situation or situations in which their participants were placed differ from those of everyday life. They should also take account of the desirability of measuring behaviour that is representative of behaviours that occur naturally. At the very least, they should interpret their findings cautiously if there are several major differences. Finally, they should discuss relevant published research that indicates the likely impact of these differences on participants' behaviour.

Ecological validity

The term ecological validity refers to the extent to which any study's findings can be generalised to other settings. Although many laboratory studies may lack ecological validity, so do some of those conducted in natural settings.

Consider Skinner's work on pigeons pecking at a disc to receive food pellets. Could the results of his study be generalised to explain how dog handlers train their dogs to seek out illegal drugs and explosives? Do the procedures for operant conditioning remain the same, i.e. the use of reinforcement to shape behaviour?

Imagine you are an observer watching birds in their natural environment, collecting data on how the parents are caring for their offspring. You disturb the parent birds by making too much noise, and they abandon their nesting site. Would your research have ecological validity because it was carried out in the natural environment? Could you generalise your findings to other settings?

WRITING UP A PRACTICAL

Practicals in psychology are written up in a standard way. Thus, your write-ups need to be organised in a certain fashion. Initially, this may seem difficult. However, it has the great advantage that this organisation makes it easy for someone reading your write-ups to know where to look to find information about the type of participants used, the statistical analyses, and so on. The details of how to produce a write-up differ slightly depending on whether it is based on an experimental or a non-experimental design. However, the general approach is exactly the same, and the essence of that approach is given later. The sections are arranged in the order they should appear in your write-ups. It is essential to refer to coursework assessment criteria issued by the relevant examination board.

Finally, be sure to write in a formal way. For example, write "It was decided to study the effects of attention on learning" rather than, "I decided to study the effects of attention on learning."

Title

This should give a short indication of the nature of your study. In the case of an experimental study, it might well refer to the independent and dependent variables. A non-experimental study would include reference to the qualitative nature of the investigation.

Abstract

This should provide a brief account of the purpose of the study, the key aspects of the design, the use of statistics, and the key findings and their interpretation.

Introduction

This should start with an account of the main concepts and background literature relevant to your study. It should then move on to a consideration of previous work that is of *direct* relevance to your study. Avoid describing several studies that are only loosely related to your study.

Aim

This resembles the experimental hypothesis, but is more general in that it indicates the background to the hypothesis.

Method

Design. Here you should indicate the number of groups, the use of an independent samples or repeated measures design (if applicable), the nature of the independent and dependent variables (if any), the experimental hypothesis, and the null hypothesis. You should also indicate any attempts made to control the situation effectively so as to produce an effective design.

Participants. The number of participants should be given together with relevant information about them (e.g. age, gender, educational background). You should indicate how they were selected for your study and, in the case of an experiment, refer to the way in which they were allocated to conditions.

Apparatus and materials. There should be a brief description of any apparatus used in the study, together with an account of any stimuli presented to the participants (e.g. 20 common 5-letter nouns). The stimuli should be referred to in a numbered section in an appendix where they can be examined in detail.

Procedure. The sequence of events experienced by the participants, including any instructions given to them, should be indicated here. Standardised instructions may be given in detail in an appendix.

Results

It is generally useful to restate the aims of the study and to indicate the independent and dependent variables in the case of an experiment.

Also, it is desirable to provide a summary table of the performance of participants. Tables of central tendency and standard deviation are usually informative ways of getting an overall “picture” of results. A bar chart or some other suitable figure may provide ready visual access to a large body of information.

- Make sure that tables and figures are clearly labelled.
- Make sure that raw data appear in a numbered section of the appendix.

Statistical test and level of significance. The test that has been applied to the data should be indicated, together with the justification for the selection of the test. Also there should be reference to the level of statistical significance that was achieved with respect to the test statistic chosen. Make sure you indicate whether a one-tailed or a two-tailed test was used, and relate your findings to the experimental and null hypotheses.

Discussion

The discussion should start by considering your findings, especially with respect to the results of the statistical test or tests. Be as precise as possible in terms of what your findings show (and do not show!). You may wish to comment on individual results that were inconsistent with the rest of the participants’ data.

The next part of this section should consist of how your findings relate to previous findings referred to in the introduction. Ask yourself if they support or refute existing theories or approaches and how you might account for the behaviour of the participants.

Next, identify any weaknesses in your study, and indicate how they could be eliminated in a subsequent study. For example, there may have been ethical issues which arose during the investigation which only became apparent after you had started.

Finally, consider whether there are interesting ways in which your study could be extended to provide more information about the phenomenon you have been investigating. This is a very satisfactory section to deal with because your imagination can take over, producing ideal studies unencumbered by the necessity to go and find participants! Always remember, though, that possible extension studies should be relevant and the likely outcome to them should be mentioned.

References

Full information about any references you have referred to in the write-up should be provided here. Textbooks (including this one) typically have a reference section set out in conventional style and you should refer to it.

PERSONAL REFLECTIONS

- Data analysis is very important. In its absence, all we could do is to interpret our data in an entirely subjective way. Data analysis has the great advantage that it allows us to be as precise as possible in our interpretations of the findings we have obtained. Data analysis sometimes seems difficult, but it is a crucial ingredient in psychological research.

SUMMARY

Qualitative research is concerned with the experiences of the participants, and with the meanings they attach to themselves and their lives. Investigators using interviews, case studies, or observations often (but not always) make use of qualitative data. A key principle of qualitative analysis is that theoretical understanding emerges from the data, and is not imposed by the researcher. Qualitative researchers typically categorise the data after taking account of all of the data and of the participants' own categories. Findings based on qualitative data tend to be unreliable and hard to replicate.

It can be hard to interpret the information obtained from interviews because of social desirability bias, complex interactional processes, and the self-fulfilling prophecy. The greatest danger with case studies is drawing very general conclusions from a single atypical individual. Case studies can suggest hypotheses, which can then be tested with larger groups. The findings of observational studies are often difficult to interpret, because it is not clear *why* the participants are behaving as they are. In addition, the participants in observational studies may not be representative.

Content analysis has been used as a method for analysing messages in the media as well as communications that participants have been asked to produce, such as diaries. The first step is the construction of coding units into which the selected information can be categorised. Coders may be asked to provide ratings or rankings as well as to categorise.

When we have obtained scores from a group of participants, we can summarise our data by working out a measure of central tendency and a measure of dispersion or spread of scores around the central tendency. The mean is the most generally useful measure of central tendency, but other measures include the median and mode. The standard deviation is the most useful measure of dispersion. Other measures include the range and the variation ratio.

Summary data from a study can be presented in the form of a figure, so that it is easy to observe general trends. Among the possible ways of presenting the data in a figure are the following: frequency polygon; histogram; and bar chart. Frequency polygons and

Qualitative analysis of data

Interpretation of interviews, case studies, and observations

*Content analysis
Descriptive statistics*

Quantitative analysis:

Data presentation

histograms are used when the scores can be ordered from low to high, whereas bar charts are used when the scores are in the form of categories.

Statistical tests

If the experimental hypothesis predicts the direction of effects, then a one-tailed test should be used. Otherwise, a two-tailed test should be used. There are four types of data of increasing levels of precision as follows: nominal; ordinal; interval; and ratio. Psychologists generally use the 5% level of statistical significance. This produces fairly small probabilities of incorrectly rejecting the null hypothesis in favour of the experimental hypothesis (Type I error) or of incorrectly retaining the null hypothesis (Type II error).

A test of difference is used when data are obtained from a study in which an independent variable was manipulated to observe its effects. The Mann-Whitney U test is the appropriate test of difference if an independent design was used. The sign test can be used when a repeated measures or matched participants design was used and the data are nominal or ordinal. The same is true of the Wilcoxon matched pairs signed ranks test, except that the data must be at least ordinal.

The data from correlational studies are in the form of scores on two response variables from every participant. These data can be presented in the form of a scattergraph or scattergram. The correlation between two sets of scores can be calculated by means of Spearman's rho test, provided that the data are at least ordinal.

The chi-squared test is a test of association. It is used when we have nominal data in the form of frequencies, and when each and every observation is independent of all the other observations. The test is nearly always one-tailed. All the expected frequencies should be five or more. Finding an association is not the same as showing the existence of a causal effect.

Issues of experimental and ecological validity

Experimental validity is based on the extent to which a given finding is genuine, and is due to the independent variable that was manipulated. A study is most likely to be high in experimental validity when all the principles of experimental design (e.g. randomisation; standardisation) have been followed. Replication provides some assurance that experimental validity is high. Ecological validity refers to the extent to which the findings of a study generalise to other locations, times, and measures. The ecological validity of a study is best assessed by carrying out a range of further studies using different locations, times, and measures.

FURTHER READING

- There is detailed but user-friendly coverage of the topics discussed in this chapter in H. Coolican (1999), *Research methods and statistics in psychology (3rd Edn.)*, London: Hodder & Stoughton. A shorter version of the Coolican textbook is H. Coolican (1996), *Introduction to research methods and statistics in psychology (2nd Edn.)*, London: Hodder & Stoughton. There is extensive coverage of the main types of qualitative analysis in P. Banister, E. Burman, I. Parker, M. Taylor, and C. Tindall (1994), *Qualitative methods in psychology: A research guide*, Buckingham, UK: Open University Press.

REVISION QUESTIONS

- 1a** What are *two* advantages of laboratory experiments over field experiments? (6 marks)
- 1b** What are *two* advantages of field experiments over laboratory experiments? (6 marks)
- 1c** What is a natural experiment? (6 marks)
- 1d** Identify *two* strengths and *two* weaknesses of natural experiments. (6 marks)
- 2** An experimenter was interested in testing the experimental hypothesis that there is an association between cigarette smoking and physical health. Accordingly, she obtained a sample of 700 people, and placed them in the following four categories: non-smoker; light smoker; moderate smoker; and heavy smoker. Their physical health was assessed as either good or poor.
- 2a** Indicate which statistical test would be appropriate to test the experimental hypothesis, and provide reasons for your choice. (9 marks)
- 2b** Describe a suitable method for obtaining the sample for the study. (6 marks)
- 2c** What is a two-tailed test? (3 marks)
- 2d** If the findings were statistically significant on a two-tailed test, what conclusions would you draw? (6 marks)
- 3** A researcher predicted that eight-year-old children would be better than six-year-old children at solving 15 simple arithmetic problems. The mean number of problems solved by the eight-year-olds was 10.8 out of 15, compared to 8.2 by the six-year-olds.
- 3a** What is meant by the term *mean*? (3 marks)
- 3b** When is it preferable to use the median rather than the mean as a measure of central tendency? (6 marks)
- 3c** Which statistical test would you use to test the prediction? Give reasons for your choice. (9 marks)
- 3d** Would you use a one-tailed or a two-tailed test, and why? (6 marks)
- 4** Research was conducted into the relationship between the amount of television violence that children watched and their level of aggression. The investigator predicted that there would be a positive relationship between the two variables. The findings from 12 children were as follows:

Participant	Television violence seen per week (minutes)	Level of aggression (maximum = 20)
1	15	8
2	24	11
3	156	18
4	29	11
5	121	10
6	84	9
7	63	7
8	68	17
9	0	5
10	58	8
11	99	12
12	112	15

The researcher used Spearman's rho to analyse the data. She obtained a value of r_s of +0.64, which is significant at $p = 0.025$ on a one-tailed test. Accordingly, she rejected the null hypothesis in favour of the experimental hypothesis that there is a positive relationship between the amount of television violence watched and the level of aggression.

- 4a** How may the data be represented? (3 marks)
- 4b** Draw a figure to represent these data. (9 marks)
- 4c** What conclusions may be drawn from your figure? (6 marks)
- 4d** Identify *two* reasons why Spearman's rho was used to analyse the data. (12 marks)
- 4e** What is a one-tailed test? (6 marks)
- 4f** What is the meaning of $p = 0.025$? (6 marks)
- 4g** What was the null hypothesis for this study? (6 marks)

REFERENCES

- Banaji, M.R., & Crowder, R.G. (1989). The bankruptcy of everyday memory. *American Psychologist*, *44*, 1185–1193.
- Bracht, G.H., & Glass, G.V. (1968). The external validity of experiments. *American Educational Research Research Journal*, *5*, 437–474.
- Brett, D.J., & Cantor, J. (1988). The portrayal of men and women in US television commercials: A recent content analysis and trends over 15 years. *Sex Roles*, *18*, 595–609.
- Cardwell, M., Clark, L., & Meldrum, C. (1996). *Psychology for A level*. London: Collins Educational.

- Coolican, H. (1994). *Research methods and statistics in psychology (2nd edn.)*. London: Hodder & Stoughton.
- Cumberbatch, G. (1990). *Television advertising and sex role stereotyping: A content analysis (working paper IV for the Broadcasting Standards Council)*. Communications Research Group, Aston University, UK.
- Freud, A., & Dann, S. (1951). An experiment in group upbringing. *Psychoanalytic Study of the Child*, 6, 127–168.
- Gould, S.J. (1982). A nation of morons. *New Scientist* (6 May 1982), 349–352.
- Jourard, S.M. (1966). An exploratory study of body-accessibility. *British Journal of Social and Clinical Psychology*, 5, 221–231.
- McAdams, D.P. (1988). *Intimacy, power, and the life history*. New York: Guilford.
- McArthur, L.Z., & Resko, B.G. (1975). The portrayal of men and women in American TV commercials. *Journal of Social Psychology*, 97, 209–220.
- Patton, M.Q. (1980). *Qualitative evaluation methods*. London: Sage.
- Reason, J.T., & Rowan, J. (Eds.) (1981). *Human enquiry: A sourcebook in new paradigm research*. Chichester: Wiley.
- Reicher, S.D., & Potter, J. (1985). Psychological theory as intergroup perspective: A comparative analysis of “scientific” and “lay” accounts of crowd events. *Human Relations*, 38, 167–189.
- Robson, C. (1994). *Experimental design and statistics in psychology (3rd edn.)*. Harmondsworth, Middlesex: Penguin.
- Schmidt, U., & Treasure, J. (1993). Getting better bit(e) by bit(e): A survival kit for sufferers of bulimia nervosa and binge eating disorders. Hove, UK: Psychology Press.
- Weiskrantz, L. (1986). *Blindsight: A case study and its implications*. Oxford: Oxford University Press.
- Whyte, W.F. (1943). *Street corner society: The social structure of an Italian slum*. Chicago: University of Chicago Press.