Part 2 Design and statistics

Part 1 of this book contains advice on how to write up your studies. The issues I describe there, together with the supplementary material you can find at **http://mcgraw-hill.co.uk/openup/harris/**, is designed to help you write better practical reports.

Part 2 of this book contains material on design and statistics. It is designed to give you the background you will need in key aspects of design and statistics to help you write better reports. For some of you this part of the book will be your introduction to these issues; for others, this section provides an opportunity to assess your understanding rather than to learn new things. Chapters 9–11 contain core background material that all of you should be familiar with before you write a report. Chapters 12 and 13 contain more advanced material that you will come to as your career as a student of psychology progresses. Make sure that you understand the material in each chapter before you move on to the next one. At the end of each chapter you will find a section designed to help you assess what you know and consolidate your learning. You can test your understanding using the diagnostic questions you will find there. You will also find there references to relevant material contained in the two statistical textbooks paired with this book – Greene and D'Oliveira's *Learning to use statistical tests in psychology* and Pallant's *SPSS survival manual* – and also, where appropriate, on the Web site that accompanies this book.

Although Part 2 has experimenting as its main focus, you will also find here coverage of relevant aspects of non-experimental approaches to quantitative research.

9 Experiments, correlation and description

The practical report described in Part 1 is designed for the reporting of *quantitative* studies in psychology – that is, studies in which you have data in the form of numbers to report. As a student of psychology you are likely to use a range of such methods, from true experiments at one end to non-experimental studies involving questionnaires, interviews, diary methods or observation at the other. In this chapter, I describe the key features of the main quantitative method discussed in this book – experimenting – and also discuss two other general approaches to quantitative research, correlation and description.

9.1 Experimenting

The first term that you must learn here is the term **variable**. In the language of design we tend to use this as a noun – that is we talk of "a variable" or "the variable". To us, a variable is quite simply something – anything – that *varies*. All around us there are variables: people come in *different* shapes and sizes, belong to *different* groups and classes, have *different* abilities and tastes. As researchers we can ask the participants in our experiments to do *different* things – to learn different lists of words for a memory test, to do boring or interesting tasks, to do push-ups for 3 minutes or to spend the same amount of time relaxing. All of these things – from social class through to whether our participants did push-ups or relaxed – are variables. So, in the above example social class is a variable because there are different social classes; doing different tasks is also a variable because

the task varies – it can be *either* interesting *or* boring. As our world is full of changing events, therefore, it is no surprise to find that almost anything in it is – or can be – a variable.

A variable, then, is something that can come in different forms or, as we say in design terms, that can take different *values* or *levels*. In practice, it can be anything from the number of white blood cells in a cubic centimetre of blood to whether or not someone voted in the last election. With our scientific hats on we are extremely interested in variables, especially in finding out something about relationships *between* them.

Our ultimate aim is to find out which variables are responsible for *causing* the events that take place around us. We see a world full of variables and we want to know what factors are responsible for producing or causing them. Why is it that some people react with depression to events that others take in their stride? Why is it that some public speakers are more persuasive than others? What makes some events easier to recall than others? We believe that we live in a universe of causes and effects and our adopted task is to try to determine, as far as is possible, which are the causes and what are the effects they produce. That is, ideally we want to be able to infer cause and effect relationships, to make what are known as **causal inferences**. Our principal weapon in this battle is the psychological experiment.

9.1.1 The experiment

The principles of basic experimenting are dead easy. In an **experiment** what we do – quite simply – is play around with one variable (the one we suspect to be the **causal variable**), and see what happens to another variable (the **effect variable**). That is, as experimenters we deliberately alter the values or levels of the causal variable – or, as we say in experimental design parlance, we **manipulate** this variable – and look to see if this produces corresponding changes in the other – the effect – variable. If it does, and we cannot see any other variable in the situation that may have produced this effect, we assume that the variable we manipulated has indeed produced the change we observed. That is, we infer a cause–effect relationship between the two variables (i.e., make a causal inference). This is the logic of experimental design.

For instance, if we suspected that eating foods that contained particular additives was responsible for causing changes in mood, we could *vary* the intake of these additives among our participants and

see whether this produced any corresponding changes in their mood. That is, we would *manipulate* the variable food additives and *measure* the variable mood. Similarly, if we wanted to find out whether physical exercise affected mental alertness, we would *manipulate* the variable physical exercise and *measure* the variable mental alertness. In the above examples, food additives and physical exercise are the variables that we suspect to be causal; mood and mental alertness (respectively) the variables we think that they influence.

However, because they are both variables, and yet play quite distinctive roles in an experiment, we give these critical variables different names. The one we play around with – the variable we manipulate – we call the **independent variable** (IV). The IV, therefore, is the variable we suspect is the *causal* variable. The variable we look at to see if there are any changes produced by our manipulation of the causal variable – that is, the variable we *measure* – we call the **dependent variable** (DV). The DV, therefore, is the variable that shows us whether there is any effect of changing the values of the independent variable. If there is such an effect, then the values that the dependent variable takes will *depend* on the values that we, as experimenters, set *independently* on our independent variable.

It is important that you get this straight. These terms are critical and you will need to apply them properly in your report. So read back through the last paragraph carefully, and then answer the following questions.

? SAQ 24

In an experiment, what is the name of the variable we *manipulate*? What is the name of the variable we *measure*?

For each of the following experiments, write down what the *independent variable* is and what the *dependent variable* is:

(a) An experimenter is interested in the effect of word frequency upon the time taken to decide whether a stimulus is a *word* or a *non-word* (a meaningless combination of vowels and consonants). She exposes the same set of participants to three sets of words which vary in their frequency (high, medium, or low) and measures the time that it takes them to decide whether they have seen a word or a non-word in milliseconds.

(b) A researcher is interested in the effect of the sex hormone oestrogen on the feeding behaviour of female rats. He injects one group with a suitable concentration of oestrogen and another group

with an equivalent volume of saline solution. After 3 days he measures the changes that have taken place in their body weights in grams.

(c) A social psychologist is interested in the role that anxiety plays in persuasion. She develops three separate public information programmes on dental care. These programmes vary in the extent to which they arouse anxiety in their viewers: one provokes a comparatively high level of anxiety, another a moderate degree of anxiety, and the third comparatively little anxiety. She exposes three separate groups of participants to these programmes and then assesses how many of those in each of the three groups subsequently take up the opportunity to make a dental appointment.

(d) An experimenter is interested in the effects of television violence upon the level of aggression it induces in its viewers. He exposes three separate groups of participants to three different kinds of television programme: one in which the violence is "realistic" (i.e., the blows cause obvious damage to the recipients), one in which the violence is "unrealistic" (i.e., the blows do not appear to damage the recipient, or hinder his ability to continue fighting), and one in which no violence is portrayed. The participants are subsequently allowed to administer electric shocks to their own victims during the course of a simulated teaching exercise (although, unbeknown to them, no actual shocks are delivered). The experimenter measures the mean level of shock (in volts) administered by those in each of the three groups.

(e) An occupational psychologist wishes to examine the impact of working with others on the productivity of a group of factory workers. She measures the number of packets of breakfast cereal this group pack into boxes in a 20-minute period when: working alone; working with one other worker; working with two others; working with four others; and working with eight others.

As experimenters we only control the values of the *independent* variable; we have no control over the precise values taken by the *dependent* variable. You will obtain these latter values from your participants – reaction times (in milliseconds), number of errors made, scores on a scale (e.g., for extraversion on a personality scale), blood glucose level (e.g., in milligrams per 100 millilitres), and so on. Note that you must *always* mention the units in which your dependent variable was measured.

Summary

1 A variable is anything that can come in different forms.

2 In experiments we manipulate variables in order to see what effects this has on other variables – in other words, which variables *cause* changes in other variables.

3 The variable we *manipulate* in an experiment is known as the *independent variable*.

4 The variable we *measure* in an experiment – to see if our manipulation of the independent variable has caused any changes – is known as the *dependent variable*.

9.1.2 Experimental and control conditions

When we manipulate an independent variable, therefore, what we do is alter the values or **levels** that it takes. We then look to see if altering these values or levels produces any corresponding changes in participants' scores on the dependent variable. Where our experiment has a single IV, the different values or levels of this variable are the *conditions* of our experiment. What we are interested in doing is *comparing* these conditions to see if there is any difference between them in the participants' scores. This, in a nutshell, is the logic of experimental design. Now let us see how it all looks in an experiment.

Suppose that you and I are interested in "folk wisdom" – and, in particular, in the old adage that eating cheese shortly before going to bed gives people nightmares. One basic way that we might test this proposition experimentally would be to take two groups of people, give one group a measured quantity of cheese (say, in proportion to their body weight) a standard time before going to bed (say, 3 hours) and ensure that those in the other group consumed no cheese during the same period. We could then count the number of nightmares reported by the two groups.

? SAQ 25 What are the independent and dependent variables in this experiment?

In this case we have manipulated our IV (cheese consumption) by forming two *conditions*: one in which the participants eat cheese

(Condition 1), and one in which they do not eat cheese (Condition 2). We are interested in the consequences of this – in the effect that this will have on the number of nightmares the participants experience in the two conditions.

This is a very basic experimental design. Our manipulation of the IV in this case involves comparing what happens when the suspected causal variable (cheese) is *present* (Condition 1) with what happens when it is *absent* (Condition 2). If cheese *does* cause nightmares, then we would expect those in Condition 1 to experience more nightmares than those in Condition 2. If cheese does *not* cause nightmares, then we should expect no such difference.

There are particular names for the conditions in an experiment where the IV is manipulated by comparing its *presence* with its *absence*. The condition in which the suspected causal variable is *present* is called the **experimental condition**. The condition in which the suspected causal variable is *absent* is called the **control condition**.

? SAQ 26 Which is the *control* condition in the above experiment?

Experiments in which we simply compare an experimental with a control condition can be simple and effective ways of finding things out in psychology. However, such a design does not exhaust the possibilities. We may, for instance, be more interested in comparing the effects of two different *levels* of the IV, than in comparing its presence with its absence. For instance, we might be interested in finding out whether different *types* of cheese produce different numbers of nightmares. We might therefore wish to compare a group of participants that ate Cheddar cheese with a group that ate Gorgonzola. In such an experiment, cheese would be present in *both* conditions. We would therefore have an experiment with two experimental conditions, rather than a control and experimental condition.

Such a design is perfectly acceptable in psychology. In fact, we do not even need to restrict ourselves to comparing only *two* experimental conditions – we can compare three or even more in one experiment. Moreover, one of these can be a *control* condition if we wish – that is, a condition in which the suspected causal variable is absent. So, for instance, we might expand our cheese and nightmare experiment to one in which we had a group of participants who ate Cheddar cheese, a group who ate Gorgonzola, a group who ate Brie, a group who ate Emmental, and a group who ate no cheese at all.

? SAQ 27 How many conditions are there *overall* in this version of the cheese experiment? How many of these are *experimental* conditions?

? SAQ 28 Go back through the experiments described in SAQ 24, and state how many conditions each independent variable has. Do any of these have a control condition? If so, which ones?

I will discuss designs involving more than two levels of the IV in Chapter 13.

9.1.3 Control: eliminating confounding variables

What we do in an experiment, therefore, is manipulate the variable that we suspect to be causal – what we have learned to call the independent variable – and examine what impact this manipulation has upon the effect variable – what we have learned to call the dependent variable. If we subsequently find that there are indeed changes in the dependent variable, this strengthens our suspicion that there is a causal link between them.

However, in order to make this causal inference we have to ensure that the IV really is the only thing that we are changing in the experiment. Thus, in our cheese and nightmare experiment, for example, we do not want more people who are suffering life crises, or taking sleeping pills, or who are otherwise prone to having nightmares, to end up in the one condition rather than the other. If they did, we would no longer be sure that any differences we observed in our DV (incidence of nightmares) would be due to our manipulation of the cheese variable, or to these other variables. For, if our groups of participants indeed varied in their intake of sleeping pills, it might be that the variable *sleeping-pill consumption* is responsible for any differences between the groups in the extent to which they experience nightmares, rather than our IV (cheese consumption).

A variable that changes along with our IV is always an alternative possible cause of any differences we find in our DV. It represents a rival explanation for these effects, and is consequently an unwelcome intruder. We must, therefore, attempt to eliminate such variables and we do this by *controlling* them.

The most effective way of achieving **control** is to hold the other main candidates for the causal variable *constant* throughout our experiment, while changing the IV alone. Thus, in our cheese and nightmare experiment we would attempt to ensure that the only important difference between our groups of participants was whether or not they ate cheese before going to bed. Consequently, we would do our best to make sure that the groups really did not differ in their sleeping-pill consumption, or life crises, or any other variable that might make them prone to experience nightmares.

An uncontrolled variable that changes along with our independent variable is known in the trade as a **confounding variable** (to confound = to confuse), for it confounds the effects of the independent variable. Confounding variables are not just any old uncontrolled variables. Confounding variables are special – they are variables that themselves have different levels that *coincide* with the different levels of the IV – things like finding that more people with sleep problems are in the experimental than in the control condition (or vice versa). Confounding variables are spanners in the works of the experiment, fundamental flaws in the design or execution that undermine the capacity to draw causal inferences.

You will find that much of the time you spend on the design of your experiments will involve spotting and controlling potential confounding variables. Confounding variables prevent us from unequivocally attributing the changes that we find in the DV to our manipulation of the IV. This is because they provide possible alternative explanations – for example, that any difference in the occurrence of nightmares is caused by sleeping pills, rather than by cheese consumption.

To recap, then, in an experiment what we do is take a variable, the variable that might be causal (the independent variable) and manipulate this variable (e.g., by varying whether it is absent or present). We then look for associated changes on the effect or outcome variable (the dependent variable). If we find that the effect is present when the suspected cause is present, but is absent when the suspected cause is absent, *and if* we have held everything else *constant* across the two conditions of our experiment, then we can conclude that the variable is indeed causal. There can be no other explanation for the changes in the effect variable, because the only thing that differs between the two conditions is the presence or absence of the causal variable. If, however, we find that the effect occurs, even though the suspected cause is not present, then we can safely conclude that it is not the causal variable after all.

Summary of Sections 9.1.2–9.1.3

1 In order to design an effective experiment we have to isolate the variables that we are interested in and hold other variables constant across our experimental conditions. This is known as *controlling* these other variables.

2 Variables that change along with (*covary* with) our independent variable are *confounding* variables.

3 A well-designed experiment does not have confounding variables.

9.1.4 Experimental and null hypotheses

In an experiment, then, we *manipulate* an independent variable by altering the *values* that it takes. Simultaneously, we *control* other candidates for the causal variable by holding them *constant* across the conditions. We then look to see if the changes we make to the IV result in changes to the DV – that is, whether there are differences in the DV between our conditions. In particular, if our IV really is the causal variable, then we can anticipate, or *hypothesize*, what the actual outcome will be. This is a fundamental feature of experimenting. It is time to consider it in some detail.

Think back to the cheese experiment. If folk wisdom is correct, then what would we expect to happen in this experiment? Wisdom has it that eating cheese shortly before going to sleep will result in nightmares. Therefore, we should expect those in the cheese condition to experience more nightmares than those in the control condition. That is, if cheese really does cause nightmares, then we would expect more of those eating cheese before going to bed to experience nightmares than do those not eating cheese before going to bed.

? SAQ 29 However, is this the only possible outcome? What else might happen? Put the book down for a moment and think about this clearly.

In fact, our experiment can have one of *three* possible outcomes. We might find that those in the experimental condition experience *more* nightmares than do those in the control condition. On the other hand, we might find that those in the experimental condition experience *fewer* nightmares than do those in the control condition. Finally,

we might find that there is essentially *no difference* between the conditions in the number of nightmares the participants experience. This will be true of *all* experiments in which we test for differences between two conditions; between them, these three options exhaust the possibilities. So – and this is important – we can specify *in advance* what the outcome of our experiment might be.

However, there is an important distinction to be made between these three potential outcomes. Two of them predict that there will be a *difference* between our conditions in the incidence of nightmares. These happen to be the outcomes we would anticipate *if* cheese affects nightmares in some way (either by stimulating or suppressing them). The third outcome, however, is that there will be no noteworthy difference between them. This is what we would expect if cheese has no influence upon nightmares.

Thus, if we assume that cheese does affect the occurrence of nightmares in some way, we would predict a different outcome to our experiment than if we assume that cheese has no effect upon the incidence of nightmares. That is, if we assume that eating cheese in some way changes the likelihood of experiencing nightmares, we would predict that there will be a difference of some sort between our groups in the frequency of nightmares they report. Of course, the assumption that cheese influences the occurrence of nightmares is the very assumption that led us to design our experiment in the first place. So, under the assumption that led to the experiment, we would predict a difference of some sort between our two groups. For this reason we call this assumption the **experimental hypothesis**. In fact:

1 *All* experiments have experimental hypotheses.
2 The experimental hypothesis leads us to predict that there will be a difference between conditions. That is, from the experimental hypothesis we derive the prediction that there will be a difference on the DV as a result of manipulating the IV.

? SAQ 30 Go back to SAQ 24 and work through the examples there, stating for each what you think the prediction might be under the experimental hypothesis.

Now, the predictions derived from the experimental hypotheses come in one of two forms. The **nondirectional prediction** states that there

will be a difference somewhere between the conditions in your experiment, but says nothing about the *direction* of this difference (i.e., does not state which condition will exceed the other on the dependent variable). In contrast, the **directional prediction** not only leads us to expect a difference, but also says something about the direction that the difference will take. (Alternative names that you might come across for these are **bidirectional** for the nondirectional prediction and **unidirectional** for the directional prediction.)

Thus, in our cheese and nightmare experiment, the nondirectional prediction simply states that the experimental and control group will experience *different* numbers of nightmares. A directional prediction, on the other hand, would state, for example, that our cheese group will experience *more* nightmares than our no cheese group.

However, what of the third possible outcome – the prediction that there will be *no difference* between the two conditions? This is a very important prediction and is derived from an hypothesis called the **null hypothesis**. You will find that the null hypothesis plays a crucial role in the process by which you come to analyse your data. In fact:

1 *All* experiments have a null hypothesis.
2 The null hypothesis is that the independent variable does not affect the dependent variable. It therefore leads us to predict that there will be little or no difference on the dependent variable between the conditions in the experiment.

Under this null hypothesis we assume, therefore, that the IV does *not* cause changes in the DV. It is called the null hypothesis because it is capable of being tested and thus of being *nullified*. As you will discover in Chapter 11, we use the null hypothesis to help us to calculate the statistics that we need to make sense of our data. Despite its name, the null hypothesis is thus a very important hypothesis.

It is very important that you learn to distinguish clearly between the predictions derived from the null hypothesis and those derived from the experimental hypothesis. The experimental hypothesis leads us to predict a difference between the conditions. In contrast, the null hypothesis leads us to predict that there will be no such difference. Do not confuse these! Design your experiments so that the *experimental* hypothesis leads to predictions of a difference *somewhere* between your conditions. Otherwise you will actually be failing to test your theoretical ideas (because you will be confusing the predictions

derived from the experimental hypothesis with those derived from the null hypothesis – see Appendix 1).

This is why it is important not to think of the experimental hypothesis as the *experimenter's* own personal hypothesis. It is perfectly conceivable that you might design and run an experiment in which you do not really expect there to be a difference between your conditions (e.g., if you are testing a theory that you do not agree with). Nevertheless, the experimental hypothesis would still lead you to predict a difference, even if this was not the hypothesis that you favoured or expected. For instance, you might be sceptical about the validity of folk wisdom and thus expect (privately) to find no difference in the incidence of nightmares in the two conditions (eating or not eating cheese). Nevertheless, the *experimental* hypothesis will still lead you to predict such a difference. For this is the proposition under test in the experiment.

Summary of Section 9.1.4

1 All experiments have a null hypothesis and at least one experimental hypothesis.

2 The experimental hypothesis involves the assumption that the IV affects the DV. It leads to the prediction that there will be a *difference* on the DV somewhere among the conditions in the experiment.

3 The prediction derived from the experimental hypothesis can either lead us to expect a difference between conditions in one direction (*directional*) or in either direction (*nondirectional*).

4 The null hypothesis involves the assumption that the IV does *not* affect the DV. It therefore leads to the prediction that there will be *no difference* among the conditions in the experiment.

9.1.5 More on controlling variables

As budding experimenters, you will come to appreciate not only that the world is full of variables, but that when you have on your experimenter's hat, most of them, most of the time, are at best irrelevant and at worst a downright nuisance. When we are interested in whether cheese causes nightmares, as well as consuming or not consuming our measured amount of Cheddar at a standard time before retiring for

the night, our experimental and control participants will also vary in what they had for their evening meal, how long they take to drop off to sleep, what they wear in bed, what colour the bedroom wallpaper is, what TV programme they watched and for how long that evening, whether they drank any alcohol, how many in-laws they have, whether and how they voted in the last election, how recently they last went swimming and whether they believe in God. Some of these variables may be obviously relevant to the issue at hand, the relationship between the IV and the DV, and others will not. For example, drinking alcohol will influence a person's quality and depth of sleep and may, therefore, influence whether or not they experience nightmares. For this reason we would not want to discover that too many more of our participants in one of the conditions drank alcohol over the course of the experiment than did participants in the other condition.

? SAQ 31 Why not?

Indeed, we may well be concerned enough about the possibility that alcohol would nullify or otherwise distort the relationship between our IV and DV that we would want to make this one of the variables that we held constant across our experimental conditions. Unpopular though it might make us, we might therefore outlaw alcohol consumption for our participants during the period in which they took part in the experiment.

However, what about everything else? Should we attempt to control for the number of in-laws? The colour of the bedroom wallpaper? Whether our participants believe in God? Well, whereas in theory we would control absolutely everything if we could, so that the *only* difference between our conditions was the IV, in reality – in human work at least – this is not possible. Instead, we use our judgement and discretion to decide which of the millions of irrelevant or **extraneous variables** (extraneous = not belonging to the matter in hand) swishing around in the psychological universe we need to control by holding them constant. Other variables we control for, not by holding them at a constant level, but by attempting to ensure that they will vary equally enough across the conditions of our experiment to not be confounding variables. We will talk more about this in the next chapter, when we consider *randomization*. For now, however, you need to recognize that you cannot control for absolutely everything. Instead, you have

to decide on the priorities for control. That is, you have to decide which of the variables in the situation you will be able to control by holding them constant across conditions.

Although this may sound pretty confusing, it is in fact relatively straightforward. With the cheese and nightmare experiment, we have already sketched a basic design: One group of people will consume a measured amount of cheese a standard time before going to bed. The control group will not eat cheese before going to bed.

? SAQ 32 It would be even better if the control group ate a standard amount of something that looked like cheese, but contained none of the ingredients of cheese, an equivalent time before going to bed. Why?

We have already mentioned some of the other key variables that we would wish to control by holding them constant: we would probably exclude from the study anyone on medication likely to influence sleep quality (thus holding this variable at the level *no relevant medication* across both conditions). We would also probably not allow alcohol consumption (thus holding this variable at the level *no alcohol* across both conditions). We might also investigate the sleep history of our participants and exclude from the study anyone with a history of sleep problems (thus holding this variable at the level *no sleep problem history* across both conditions). In doing this you can see that, if we do find more nightmares in our experimental condition, we can rule out explanations that these are caused by more people in the experimental group having their sleep disturbed by medication, alcohol, or just being more prone to sleep problems in the first place.

There may be other variables that we would control for in this way, but we would have to stop somewhere. For instance, our participants will sleep in rooms with different coloured walls. We would be unlikely to demand that for the purposes of our experiment they all had their rooms painted a standard shade of a restful pastel colour. Thus, participants in both conditions will sleep in rooms of different colours and the variable *colour of bedroom walls* will be left to vary rather than being held constant. However, we would hope that there was no consistent or *systematic* difference in the colours in the bedrooms of the experimental group when compared with the control group.

Of course, if we were very concerned that an environmental factor like the colour of the bedroom walls was likely to make a difference,

then we might well take active steps to control for it. One way in which we could rule out such environmental factors would be to have the experiment take place not at home but in a sleep laboratory. There we could ensure that a whole range of additional factors was held constant across our experimental and control conditions. Laboratories offer opportunities for controlling a whole range of extraneous variables, which is one of the reasons for their appeal in psychology.

Summary of Section 9.1.5

1 In reality, there are usually too many irrelevant or *extraneous* variables in the situation for us to hold all of them constant.

2 We thus select the more important of these to be controlled by holding them constant, so that each participant has the same value or score on this variable.

3 The remaining variables we allow to vary among the participants *within* our conditions, but try to ensure that the overall or average scores or values for this variable do not differ *systematically* between conditions (i.e., that if calculated, the average score would be roughly equal for each of the conditions).

4 However, this control at the level of the group or condition, rather than at the level of the individual participant, does not *guarantee* that there will be no systematic differences between the conditions.

9.2 Correlation

I have more to say about designing experiments in Chapters 10 and 13. However, it is not always possible to conduct experiments. There may, for instance, be practical or ethical factors that prevent us from actually *manipulating* variables in the way we need for our studies to qualify as true experiments. Under such circumstances we can turn to a rather less powerful, but still informative, technique – **correlation**.

The critical difference between the experiment and correlation is that, with correlation, we are unable to distinguish between *independent* and *dependent* variables. This is because we do not *manipulate* variables when we undertake correlational work. Instead, we rely on *natural* changes or differences to tell us something about which variables are related in some way to each other.

To illustrate this, imagine that you and I are interested in whether living close to the transmitters that relay messages to and from mobile (cell) phones is a cause of behavioural problems in children. An experimental approach to this question would involve assigning people from birth randomly to live either close to or far away from such transmitters and then comparing the numbers in each group that eventually developed behavioural problems. This is not only profoundly unethical but clearly also impractical. Under these circumstances, if we wish to gather data on humans, then we are restricted to *correlational* data. In this instance we rely on differences that exist already in the population. That is, we would look to see if there was any correspondence between the distance existing groups of people lived from the transmitters and the incidence of behavioural problems. However, this has significant implications for our ability to say whether there is a *causal* relationship between the variables "distance from transmitter" and "behavioural problems".

For instance, suppose that we find that there is a relationship between these variables and that this is indeed the one we suspect: that, as the transmitters get closer, so there is an increase in the incidence of behavioural problems. This would be an example of what we call a **positive correlation** (because *increases* in proximity to the transmitters are accompanied by *increases* in the incidence of behavioural problems). Such a relationship is depicted graphically in Figure 9.1.

Now, at first glance it may seem obvious to conclude that the transmitters are responsible for *causing* the behavioural problems. However, we actually have no grounds for drawing this conclusion. For, with correlational data, all we know is that there is an *association* between these variables – that as one varies (goes up, goes down), so does the other. It could be that the transmitters do indeed cause the problems. On the other hand, it could be that people with behavioural problems end up living close to transmitters (perhaps because they have less control over where they live). It could also be that something else is responsible for the relationship. Perhaps the transmitters are placed in poorer areas of towns and cities, where people need the money from the telephone companies or feel too powerless to object to them placing transmitters in their communities. In which case the association may not be between the transmitters and the behavioural problems, but a **third variable**, such as poor diet or unsatisfactory schooling, may be responsible for the behavioural problems and the relationship between transmitters and behavioural problems is more apparent than real. Such a relationship is described in the trade as **spurious** (spurious = not genuine).

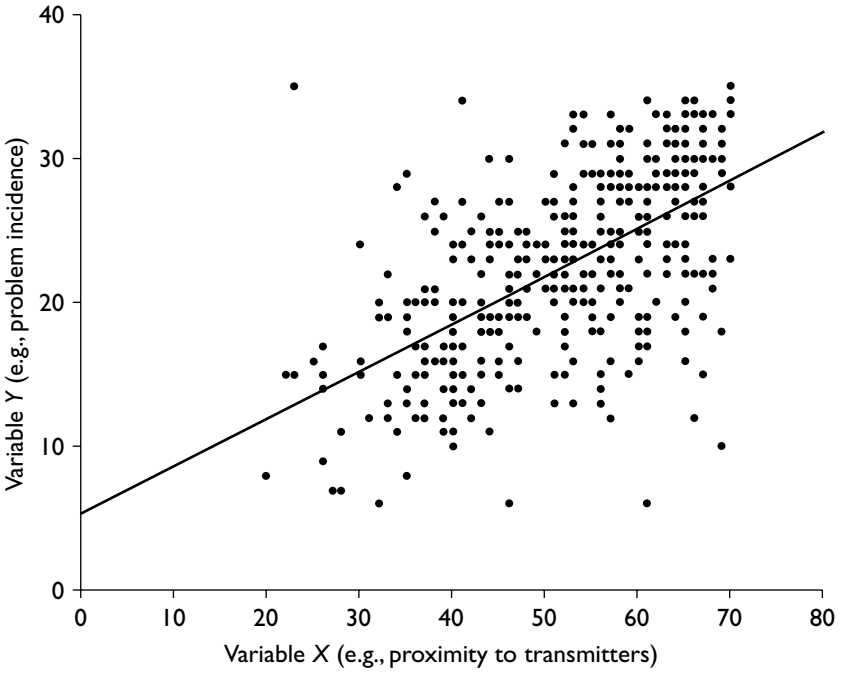

Figure 9.1. A scatter plot showing a positive correlation between variables X and Y.

An illustration might reveal this more clearly. There is, in fact, a positive correlation between the consumption of ice cream and the incidence of deaths by drowning. That is, as people tend to eat more ice cream, so more people tend to die in drowning accidents.

? SAQ 33 Does this mean that there must be some kind of causal relationship between them (even if we cannot say which way it goes)?

Now the relationship between ice cream consumption and deaths by drowning is sufficiently bizarre for us to suspect that something strange is going on. Nevertheless, *even where the direction of causality seems obvious* it is still not permissible to assume that one of the variables is responsible for producing changes in the other. For, in relying upon natural changes, rather than manipulating variables, we are unable to control variables that *covary* with (i.e., vary along with) the ones that we are interested in. That is, we are unable to

control for *confounding* variables (for, as you know, a confounding variable is an extraneous variable that covaries with the variable we are investigating).

Correlational data are thus *inevitably* confounded. So, be aware of this problem – it applies to many issues within and outside psychology, even those – such as smoking and lung cancer, or the consumption of saturated fats and heart disease – in which the direction of causality seems "obvious". With such data we cannot rule out the possibility that the causal relationship is the reverse of the one we believe or even spurious. (Thus, although the experimental data on animals and other data suggest it is highly probable that smoking causes lung cancer and excessive consumption of saturated fat contributes to heart disease, it remains possible that those who are prone to lung cancer are those who for some reason choose to smoke, or that those who are prone to heart disease are those who for some reason tend also to like eating saturated fat.)

As you can see, it is no surprise that the biggest debates in our scientific and public lives often centre around issues for which the data, concerning humans at least, are *correlational*. Moreover, as you can see from the above examples, this is certainly not an issue that is restricted to psychology.

9.3 Description

Rather than concerning ourselves with relationships between variables, we may be content simply to describe particular variables. For instance, we might wish to examine the nature of people's attitudes towards genetically modified crops, or to find out what they can tell us about some aspect of their social behaviour. Such studies are neither experimental nor correlational. They are *descriptive*.

A classic example of such an approach is the public opinion poll. Opinion polls tell us things like the percentage of the sample questioned who stated that they would be prepared to vote for particular parties if an election were held immediately, or who they think would make the best prime minister or president. This is **description**. It is simply an attempt to make a statement about the characteristics of the variable "the voting public". However, there is no attempt to *explain* these findings. That is, the data themselves are not used to explain *why* the individuals sampled intend to vote the way they do. Data of this nature are usually generated by *questionnaire* or

interview, take the form of what is called a *survey*, and are described solely in terms of *descriptive statistics*, such as percentages and means (Section 4.1). It is highly unlikely that as a student of psychology you will generate numerical data that are simply descriptive. If not experimental, your quantitative data are likely to be correlational.

Summary of Sections 9.2–9.3

1 The critical difference between the correlation and the experiment is that in the correlation we are unable to distinguish between *independent* and *dependent* variables. This is because we do not *manipulate* any variables in a correlation. Instead, we rely upon differences that already exist to discover which variables are related to each other.

2 The consequence of this is that we are unable to draw *causal inferences* from correlational data. That is, correlations reveal *associations* between variables, rather than causes and effects.

Consolidating your learning

This closing section is designed to help you check your understanding and determine what you might need to study further. You can use the list of methodological and statistical concepts to check which ones you are happy with and which you may need to double-check. You can use this list to find the relevant terms in the chapter and also in other textbooks of design or statistics. The diagnostic questions provide a further test of your understanding. Finally, there is a description of what is covered in the two statistics textbooks paired with this book plus a list of relevant further material that you can find on the book's Web site.

Methodological and statistical concepts covered in this chapter

In this chapter we discussed the following issues. You should now understand what is meant by the following terms:

Causal inference
Causal variable
Confounding variable
Control condition
Controlling extraneous variables
Correlation (positive, negative)
Dependent variable
Description
Directional (bidirectional, unidirectional) hypothesis
Effect variable
Experiment
Experimental condition
Experimental hypothesis
Extraneous variables
Independent variable
Levels of a manipulated variable
Manipulating variable
Non-directional hypothesis
Null hypothesis
Placebo
Self-fulfilling beliefs
Spurious correlation
Third variable
Variable

Diagnostic questions – check your understanding

1 What is a *variable*?
2 What is the name of the variable we *manipulate* in an experiment?
3 What is the name of the variable we *measure* in an experiment to assess the effects of our manipulation?
4 From which hypothesis can we predict that there will be *no difference* between the conditions in an experiment? Is it (a) the experimental hypothesis or (b) the null hypothesis?
5 What is a *nondirectional* prediction?

You can find the answers at the rear of the book (p. 275).

Statistics textbooks

∑ The books paired with *Designing and Reporting Experiments* have the following coverage:

Learning to use statistical tests in psychology – Greene and D'Oliveira
Greene and D'Oliveira cover variability in Chapter 1, and independent and dependent variables, experimental and null hypotheses, and experimental and control conditions in Chapter 2. They also cover different tests for correlation in Part IV. As an undergraduate student you will only need to look in Part IV if you need to find out about a particular test and you are likely at first to need the earlier chapters in this section (on Chi-square, Pearson's *r*, and simple regression) rather than the later ones, (which cover more advanced tests and issues (e.g., multiple regression and the General Linear Model).

SPSS survival manual – Pallant
Part four of Pallant covers correlation. You need turn to these chapters only once you have some data to analyse. For more advanced students, the chapter on correlation contains a useful account of how to test whether two correlation coefficients differ significantly (Chapter 11). Pallant also covers partial correlation (Chapter 12), multiple regression (Chapters 13 and 14), and factor analysis (Chapter 15). The latter are again more relevant to more advanced undergraduates and graduate students.

The *Designing and Reporting Experiments* Web site

In Section B of the Web site at **http://mcgraw-hill.co.uk/openup/harris/** you will find a discussion of how to report a wide range of statistical tests that you might use to analyse experimental or correlational data, including particular things to watch out for when you do.

10 Basic experimental design

Although experiments have a universal *basic* logic, they vary in design and complexity. For instance, they may have one or more than one independent variable, any or all of which may employ *unrelated* or *related* samples. It is time to find out what this means.

10.1 Unrelated and related samples independent variables

Suppose that you and I have been approached to undertake some research into the effect of listening to music in the car on driving performance. There are lots of ways that we could go about doing this. For instance, we might examine actual driving performance on the road, or on a test track, or perhaps we could gain access to a driving simulator and test people's performance on that. We might decide to focus on particular aspects of driving performance, such as maintaining control of the car or average speed. Or we could assess the number of mistakes made over a standard course, or monitor the participants' eye movements as they drive to discover whether there are any differences in the way that they monitor the situation unfolding outside as a consequence of having music on or off.

On the other hand, we might wish to refine the question that we address – examine whether different materials (e.g., different types of music or different types of radio programme) affect performance in different ways or to a different extent. Similarly, we might examine whether the volume at which people listen is important and whether this differs with the material being listened to. Or we might want to

find out whether teenagers are affected more than older drivers, or men more than women.

All of these possibilities hold different implications for the experiment we would eventually design and the conclusions we would ultimately draw. Essentially, we have here a whole series of decisions that we have to make about the IV(s) and DV(s) in our experiment. However, there is one fundamental feature of our design that we would still have to decide upon, even after we have determined the precise question we wish to examine, chosen our IV and DV, and set the levels on the IV. This is how we will *distribute the participants* in our experiment.

For instance, suppose we decided to run the above experiment on a driving simulator, using a computer to display a standard set of hazards and problems in a random sequence. We could then measure driving performance on a range of *relevant* measures from average speed through manual control to number of (virtual!) pedestrians knocked over. Participants would be free to choose their own listening material and to set the volume and adjust it whenever they desired.

So far, so good. However, we have yet to decide how we are going to distribute our participants across the different conditions (music on or music off). We could, for example, have one group of participants performing in the music on condition and another (different) group in the music off condition. If we did this, we would have different groups of people performing in the two conditions. The scores on our DV would therefore be from *different* samples and be *unrelated* to each other. Our IV would thus be an **unrelated samples** IV.

However, instead of assessing the driving performance of different participants, we could compare the driving performance of the *same* participants. That is, we could measure each participant's performance when driving with music off and his or her performance when driving with music on. In this experiment we have the same group of participants performing in the two conditions. The scores on our DV would therefore be from the *same* sample and would be *related* to each other. Our IV would thus be a **related samples** IV.

? SAQ 34 What type of IV did we have in our original, two-condition cheese experiment (Chapter 9)? How else could we have run this experiment, and what type of IV would we have then employed?

These, then, are the two basic IVs used by psychologists: the *unrelated* samples IV, in which we take a different batch of participants for each of the conditions or levels on the IV and the *related* samples IV, in which we use the same batch of participants for the conditions or levels on that IV.

? SAQ 35 Now go back through the experiments outlined in SAQ 24 and state what type of IVs the experimenters employed in each.

Designs involving these IVs are illustrated in Tables 10.1 and 10.2. You can see that for our experiment involving unrelated samples we need 32 participants to have 16 participants per condition. For the related samples version, however, we need only 16 participants to achieve this.

Table 10.1
An Unrelated Samples IV with Two Conditions: Allocation of Participants

Condition 1	Condition 2
p02	p01
p04	p03
p09	p05
p10	p06
p11	p07
p13	p08
p14	p12
p19	p15
p21	p16
p24	p17
p25	p18
p26	p20
p28	p22
p29	p23
p30	p27
p32	p31

Note. Participants have been allocated to conditions randomly. p = participant. Note that you would probably want to use more participants than 16 per condition (see Section 10.6).

Table 10.2
A Related Samples IV with Two Conditions

Participant	Condition 1	Condition 2
p01	2nd	1st
p02	1st	2nd
p03	1st	2nd
p04	2nd	1st
p05	2nd	1st
p06	1st	2nd
p07	2nd	1st
p08	1st	2nd
p09	2nd	1st
p10	1st	2nd
p11	2nd	1st
p12	2nd	1st
p13	1st	2nd
p14	2nd	1st
p15	1st	2nd
p16	1st	2nd

Note. Orders have been counterbalanced and allocated randomly. p = participant.

Those of you who are not expected to contribute to the design of your experiment will still need to be able to state which type of IV(s) you employed. So you must learn to make the above distinction. However, those of you who are expected to contribute at the design stage will also need to be able to decide which type to employ. So let us turn to a consideration of the factors to bear in mind when deciding which types of IV to use in your experiment. First, however, we need a brief word about terminology.

Summary of Section 10.1

1 Two basic independent variables employed in psychology are the *unrelated* samples IV and the *related* samples IV.

2 In the unrelated samples IV, a *different* set of participants is assigned to the conditions or levels of the IV.

3 In the related samples IV, the *same* set of participants appears in the conditions or levels of the IV.

10.2 Other names for unrelated and related samples independent variables

Sometimes when you are learning about methodology and statistics in psychology you can feel that you are spending almost as much time learning terminology as you are learning concepts. Moreover, just when you get to grips with one term for something you find that there is another term for the same thing (perhaps more than one additional term, which is about the last thing that you need). Sadly, this is especially true of names for this basic distinction between types of IV.

Some time ago psychologists referred to the participants in their experiments as *subjects*. These days this term is considered unacceptable, as our participants are above all our fellow human beings and not simply "subjects" of experimental enquiry. However, some important terms in statistics and methodology retain the old-fashioned language and you may well come across this terminology in some other textbooks or certain statistical software packages.

A traditional term for unrelated samples IVs is **between subjects**. This is because, when we use *unrelated* samples, our comparisons are *between* different participants. Other terms that you may come across for the same thing are **independent** or **uncorrelated samples**, **independent** or **uncorrelated measures** or **between participants** or **between groups**.

A traditional term for related samples IVs is **within subjects**. This is because, when we use *related* samples, our comparisons are *within* the same participants. (That is, we are comparing the performance of the same person across the different conditions and thus making comparisons *within* the person.) Other terms that you may come across for the same thing are **dependent** or **correlated samples**, **dependent** or **correlated measures**, **repeated measures** or **within participants**.

Unfortunately, the problem is compounded by the fact that people can become surprisingly worked up about which label should be used. Throughout this book I use "unrelated samples" to denote IVs where the data in the different levels come from different participants. I use "related samples" to denote IVs where the data in the different levels come from the same participants. Given the strength of feeling that this issue can evoke, however, it is a good idea to find out whether your tutor favours other labels for these IVs and, if so, to use these instead. The important thing (of course) is to make sure that you can

recognize these different types of IV and understand the distinction between them. What you choose to call them is secondary. However, you *must* be consistent in your use of terminology in a report.

Summary of Section 10.2

1 There are a number of different terms available to make the same basic distinction between IVs where the data in the different levels come from different participants and IVs where the data in the different levels come from the same participants.

2 The important thing is to understand the distinction underlying these types of IV and to be able to recognize which is which.

3 Do not chop and change terminology in a report: decide which terms you are going to use and stick to them throughout.

10.3 Deciding whether to use related or unrelated samples

In order to consider how you might choose between these types of IV, we must think back to the logic of experimental design discussed in Chapter 9. The aim is to manipulate the variable we suspect to be causal (the IV) and examine what impact this manipulation has upon the effect variable (the DV). If we find that, when we alter the IV, there are changes in the DV as well, this strengthens our suspicion that there is a causal link between them.

However, as we pointed out earlier, in order to make this causal inference we have to ensure that the IV really is the only thing we are changing in the situation – that is, we do not want any *confounding* variables. One source of confounding can be our particular sample of participants. Thus, in our driving and music experiment, for example, we do not want too many more of the better drivers to end up in one condition than the other. For, if they did, we could no longer be sure that any differences that we observed in our DV (e.g., number of errors made over a standard course) would be due to our manipulation of the music variable or to basic differences in driving ability. Confounding variables, remember, prevent us from unequivocally attributing the changes we find in the DV to our manipulation of the IV. This is because they provide possible rival explanations for any effects we find (e.g., that any difference in driving performance

between the conditions is due to basic differences in ability to drive rather than to the effect of having music on or off). Essentially, related and unrelated samples IVs differ in the nature and extent of their potentially confounding variables.

Let's illustrate this with an example. Suppose we ran our driving and music experiment using *unrelated* samples and found that those who drove without music performed better than those who drove with music on. This could of course be because listening to music affects concentration and impairs ability to drive. However, it could also be simply because more of the better drivers were put into the "music off" condition. For there are pronounced *individual differences* in driving ability. Any differences in driving performance between groups of different people, therefore, might well stem from this basic fact of life, rather than from any manipulation of the independent variable. When we use *unrelated* samples, of course, we are inevitably comparing the performances of different people.

There are ways in which we can attempt to minimize the impact of individual differences upon our experiment. However, there is only one way in which we can actually eliminate this source of extraneous variation. This is by employing *related* samples.

When we use related samples, we do not compare the performances of different people. Instead, we compare the performances of the *same* people on different occasions. So, in the case of our music experiment, we would assess the driving performances of any given individual both when driving with music on and when driving with music off. We can safely assume that any differences in performance under the two conditions cannot stem from individual differences in driving ability because it is the same person in both conditions.

If you are still confused by this, imagine that we developed a revolutionary kind of running shoe that we suspected would improve the running performance of those who wore it. Using *unrelated* samples would be like giving the shoes to half of the runners in a race and comparing their finishing positions with the runners who did not have the new shoes. Using *related* samples would be like getting the runners to run at least once with the new shoes and at least once with their normal running shoes. In the first race it is possible that our runners with the new shoes might find themselves up against people who were simply much better than them anyway. In the latter race, however, you can see that this would not matter, because we are not interested in where our runners came in *absolute* terms, but simply in

how well they fared in comparison with their performance without the new shoes (e.g., in time taken to complete the course). That is, in the first instance it would make a great deal of difference if they found themselves running against a team of Olympic athletes; in the latter case it would not matter at all (provided, of course, that they ran against the same athletes on both occasions and that they were not too demoralized by the first drubbing).

Using related samples, therefore, eliminates from the experimental equation permanent or chronic **individual differences** in ability at the experimental task – the things that make people *generally* better or worse at the task regardless of the IV. We know that some people will simply be better than others on the driving simulator regardless of whether they are listening to music or not. Using unrelated samples, we trust that the variation due to such individual differences in ability will be spread roughly equally between the two conditions. With related samples, however, differences in performance arising from such individual differences in ability are constant across the two conditions and can be removed from the equation.

In eliminating chronic individual differences from our experiment, related samples IVs reduce the amount of background or **extraneous variation** that we have to cope with – variation *other than* that arising from our manipulation of the IV. From our point of view, this is a good thing. For this reason, using related samples is, at least in principle, to be preferred to using unrelated samples. As you have to start somewhere, therefore, a good rule of thumb when designing experiments is to start by exploring the possibility of using related samples IVs, and only turn to the alternatives when you discover obstacles that prevent you from doing so.

Summary of Section 10.3

1 In all experiments the effects of the IV on the DV are assessed against a background of *inherent* or *extraneous* variation.

2 Permanent or chronic *individual differences* in ability at the experimental task are a considerable source of such extraneous variation.

3 Related samples IVs eliminate such *individual differences* and thereby remove a large part of this extraneous variation. Related samples IVs should, therefore, be used in preference to unrelated samples IVs when there are no insurmountable obstacles to using them.

10.4 Related samples

Related samples are preferable *in principle* to unrelated samples. However, this does not mean that using related samples is without problems.

10.4.1 Advantages

The major advantage of using related samples is that doing so reduces the background variation against which we have to assess the impact of the IV on the DV. It does this by *eliminating permanent* or *chronic individual differences*. There is also a practical advantage that, as students, you will probably find extremely useful – you need fewer participants.

10.4.2 Disadvantages

If using related samples eliminates one source of variation, it unfortunately introduces another – *order* effects. When we run an experiment using related samples, by definition, we will obtain more than one score from our participant (see Table 10.2). This means, of course, that we will have to present our conditions one after the other. This produces problems of its own – problems that do not exist when we use unrelated samples.

For example, suppose that we ran our music experiment using related samples, testing all of our participants first with the music off and then with the music on. Suppose that we found that driving actually improved when the drivers had the music on. Can you think of an alternative explanation for this finding, other than the suggestion that listening to music led to the improvement in performance?

Well, the improvement in driving performance might simply be due to the fact that the participants in our experiment got better at the task as they became more familiar with the driving simulator. That is, they might have got better because they became more *practised* and, because they all did the music on condition second, this **practice effect** contributed more to performance with the music on than to performance with the music off. In other words, we have here a *confounding* variable arising from the *order* in which we ran the conditions.

Moreover, the same argument would apply even if we had found the opposite: that is, if we had found that the performance of our drivers had *deteriorated* from the first condition to the second. In this case, it could simply be that our participants had become *bored* with the task or were *fatigued* and started to make more mistakes because of this.

Order effects are the price we pay for eliminating individual differences with related samples. Order effects come in two forms. There are those that lead to an *improvement* in the participant's performance – things like practice, increasing familiarity with the experimental task and equipment, increasing awareness of the task demands. In contrast, there are those that lead to a *deterioration* in performance – things like loss of concentration, due to fatigue or boredom. Both types of effect occur simultaneously, but have opposite effects on the DV. Both types of effect need to be controlled for.

10.4.3 Controlling for order effects

The best way of controlling for order effects is to **counterbalance**. When we counterbalance we ensure that each condition in our experiment follows and is preceded by every other condition an equal number of times. Thus, for each participant who does one particular sequence of conditions, there are other participants who perform the conditions in all the other possible combinations of orders. Although in the abstract this sounds horrendously complicated, generally it can be achieved comparatively easily. So, for instance, in our driving experiment a very simple control for order effects would be to ensure that half of the participants drove with the music on *before* they drove with the music off, whereas the other half drove with the music on *after* they had driven with the music off (see Table 10.3). This way, although we would not have *eliminated* order effects – our participants are still likely to get better or worse as time goes by – we will have rendered these effects *unsystematic*. That is, practice, fatigue and boredom should affect the music on condition about as much as they affect the music off condition.

We will talk more about counterbalancing when we consider extending the number of levels on an IV (Section 13.1.2). Sometimes, however, there are too many conditions to counterbalance. Under these circumstances there are a number of alternatives. One of these is to **randomize** the order of the conditions.

Table 10.3
A Counterbalanced Design Using Related Samples to Examine the Effects of Having Music Off or On upon Driving Performance

Participant	Music on	Music off
p02	1st	2nd
p03	1st	2nd
p06	1st	2nd
p08	1st	2nd
p10	1st	2nd
p13	1st	2nd
p15	1st	2nd
p16	1st	2nd
p01	2nd	1st
p04	2nd	1st
p05	2nd	1st
p07	2nd	1st
p09	2nd	1st
p11	2nd	1st
p12	2nd	1st
p14	2nd	1st

Note. Participants have been randomly allocated to orders. p = participant.

For example, if we had an experiment with six conditions, this would give us 720 different orders. If we wished to counterbalance these orders, we would need a minimum of 720 participants (one for each different sequence of the six conditions). Although there are alternatives that you may come across later on in your statistics course (e.g., the *Latin square*), under these circumstances you will most probably find yourselves *randomizing* the orders undertaken by your participants.

When we randomize the order of our conditions we do not actually ensure that all the conditions are followed by and preceded by every other condition an equal number of times. Instead, we trust that a *random sequence* will spread the order effects more or less equally around the various conditions. So, for instance, if we randomized the orders in which our participants performed in an experiment with six experimental conditions, we would hope that each of the conditions would appear in each of the ordinal positions (first, second, third, etc.) just about as often as the others. The more participants we have, the likelier it is that this will be the case.

This is because the critical feature of a random sequence is that the items in the sequence all have an equal chance of being selected for any of the positions in that sequence. In the case of orders, what this means is that any of the conditions can appear first, second, third, etc., for any of the participants. So, for example, in a six-condition experiment, Condition A has the same chance as Conditions B, C, D, E, and F, of being the first condition undertaken by Participant 1. If Condition C is the one actually chosen by the random process, then Conditions A, B, D, E, and F, all have an equal chance of being the second condition undertaken by Participant 1. (See Appendix 2 for the details of how to go about making such random allocations.) If Condition D is chosen as the second condition, then Conditions A, B, E, and F, all have an equal chance of being the third condition undertaken by Participant 1, and so on until all the conditions have been assigned to this participant by the random process. Moreover, the same applies to Participant 2, and indeed to *all* of the participants in the experiment (Table 10.4). (For the record, this is randomization *without* replacement – see Appendix 2.)

When used to control for order effects, however, both counterbalancing and randomization have one thing in common. They are based upon the assumption that our participants are *not much more* fatigued or bored or practised when the music off condition

Table 10.4
Randomized Orders of Conditions for the First 10 Participants in a Six-Condition Experiment Employing Related Samples

Participant	Sequence
p01	C D A E F B
p02	B D F A C E
p03	F D B A E C
p04	A C B F D E
p05	F A D E B C
p06	A B F C E D
p07	E D B F A C
p08	C D B F E A
p09	D A E B F C
p10	E F C D B A

Note. Each sequence of conditions has been created randomly and separately for each participant. p = participant.

comes after the music on condition, than they are when the music off condition comes first. If they are, then we have what we call a significant **carry-over effect** among the conditions in our experiment. Neither counterbalancing nor randomization can control for such carry-over effects.

When you suspect that the conditions of your experiment will have a carry-over effect, therefore, you should not use a related samples design. An extreme example of such a case would be, for instance, if we were attempting to compare two different techniques of teaching a particular task. Once participants have learned the task once, it is not possible to make them unlearn it in order to learn it once again by the different method. Similarly, if we were undertaking research into the effects of alcohol upon risk perception, it would be a mistake to have our no alcohol condition immediately following our alcohol condition. So you must watch out for treatment conditions in your experiments that tend to markedly alter the state of your participant. Such conditions will have a lingering effect and will consequently influence performance in subsequent conditions, thus having a carry-over effect.

If these effects are only temporary, one way around them is to introduce a longer than usual time delay between conditions (e.g., hours, days, or even weeks). However, if the effects are more or less permanent (e.g., teaching methods) or if this strategy is not feasible (e.g., lack of time, lack of co-operation on the part of your participants), then it would be better to employ one of the alternative designs.

So, when designing any experiment involving related samples IVs, think sensibly and carefully about whether there are likely to be any significant carry-over effects between conditions and, if there are, about whether it is possible to modify your design to cope with them. If not, then you need to abandon the idea of using related samples on the offending IV.

Another problem that arises with related samples is the need to duplicate and *match* materials. For example, suppose that you and I were interested in establishing whether scores on a test of reasoning are influenced by the ways in which the questions are phrased – in particular, whether the problems are couched in *abstract* terms (e.g., using algebraic expressions like $A > R$) or in *concrete* terms (e.g., by using examples drawn from everyday life to express the same relationships – things like "Antony is older than Richard"). There are profound individual differences in people's ability to reason, so ideally we would like to run this experiment using related samples. We could do this by giving participants *both* concrete and abstract

Table 10.5
A Design Using Related Samples with Matched Materials to Examine the Effects of the Phrasing of the Questions (Concrete or Abstract) on Reasoning Performance

Group	Test A	Test B
1	Concrete	Abstract
2	Abstract	Concrete

problems and comparing their performances on the two types of item. However, we would need to make sure that any differences in performance between the two conditions arose from the ways in which the problems were expressed and not from the fact that one set of problems was simply easier to solve. That is, we would need to *match* our materials so that they were equivalent in all respects other than the IV (concrete or abstract problems). In this case, this is not all that difficult to do. If you imagine two sets of problems, Test A and Test B, we can give one group of participants Test A expressed in concrete terms and Test B expressed in abstract terms, and the other group of participants Test A expressed in abstract terms and Test B expressed in concrete terms (Table 10.5). Thus, any differences we find in overall performance on the concrete and abstract items cannot be due to differences in the ease of the items themselves – because the two sets of items have appeared equally as often under the concrete and the abstract conditions.

Summary of Section 10.4

1 The cost of eliminating individual differences when we use related samples is the introduction of another source of extraneous variation – *order effects*.

2 Order effects are of two kinds: those that lead to an *improvement* and those that lead to a *deterioration* in the participant's performance on the experimental task.

3 These order effects must be *controlled* for – for example by *counterbalancing* or by *randomizing* the order of the conditions.

4 These methods do not *eliminate* the variation introduced by order. They *transform* it into unsystematic variation.

5 These methods will not work when there are significant *carry-over* effects. Under these circumstances, you should not use related samples.

6 With related samples you are likely to need to duplicate and *match* the materials.

10.5 Principal alternatives to related samples

Where there are insurmountable order effects, or it is difficult to match materials, or when the participants *have* to be different (e.g. in personality research, studies involving differences in intelligence, culture, gender, etc.), then related samples are not suitable. Under these circumstances, you should turn to one of the alternatives. The principal alternative is to use unrelated samples. However, you might also consider matching your participants in some ways (Section 10.7). If you have more than one IV, then you may be able to use a combination of related and unrelated samples IVs (see Chapter 13).

10.6 Unrelated samples

The principal alternative to the use of related samples is to use unrelated samples on your IV. As pointed out earlier, however, the biggest disadvantage of unrelated samples is the presence of *individual differences*. However, there are ways in which we can attempt to minimize their impact on our experiment.

10.6.1 Advantages

It just so happens that the advantages of unrelated samples correspond to the weaknesses of related samples. That is, there are no problems with order effects and we do not need to duplicate and match our materials.

10.6.2 Disadvantages

However, the reverse is also true. If there are no problems with order effects, this is more than offset by the intrusion of individual

differences. Likewise, if we do not have problems with materials, we have to find a lot more participants.

10.6.3 Ways around these disadvantages

With respect to the bigger disadvantage, individual differences, we must attempt to rule out any *systematic bias* stemming from such differences between participants. We cannot eliminate individual differences; we can, however, try to ensure that they are equally distributed across conditions (e.g., that as many good drivers are assigned to the music on as to the music off groups). One possibility would be to try to *match* our participants, assessing them on driving performance and making sure that equal numbers of good and bad drivers are assigned to each group (Section 10.7). In order to do this we would need to assess the performances of our participants *prior* to running the experiment (i.e., run a *pretest*).

Often, however, we have neither the resources nor the time to collect and act upon this information. Under these circumstances, the alternative we turn to is again *randomization* – in this case the *random assignment* of participants to their conditions (see Appendix 2). To do this, we go to our pool of participants and assign them randomly to conditions, trusting the random sequence to spread individuals who differ in basic ability at the experimental task more or less equally among the conditions. So, for instance, in our music and driving experiment we might assign our participants to either the music on or the music off condition using random number tables (Appendix 2). We would hope that this procedure would give us a reasonably even split of good, poor, and average drivers between the conditions. The more participants we use, the more effective this procedure is likely to be.

Random assignment of participants to the conditions of an unrelated samples IV (or to the different orders of conditions on a related samples IV) is the critical feature that makes an experiment a true experiment. Random *assignment* of participants to conditions should not, however, be confused with the random *selection* of participants from a population. Ideally you should do both. In practice, as students, you will probably only do the former (see Sections 10.8 and 10.9). Moreover, unless you check, you can never really be sure that your randomization has been effective. That is, you will be unable to state categorically that differences in ability among your participants

did *not* lead to the differences that you observed on your DV. The same is true of randomizing to control for order effects. It may be that there were systematic differences in the orders in which your conditions appeared, even though allocation of orders was randomized. So, you should bear this in mind when you come to interpret your findings. (Once you have become experienced enough, you should consider checking the effectiveness of your randomization. You can find advice on this in Section G of the book's Web site at **http://mcgraw-hill.co.uk/openup/harris/**)

With regard to the problem of having to obtain larger numbers of participants with unrelated samples than with the related samples equivalent, like the problems with materials in related samples designs, this is simply something that you have to live with. However, where you have more than one IV you may be able to reduce the number of these that require different participants by employing what we call a *mixed* design (Section 13.3) and thereby reduce the number of participants that you need overall.

Summary of Sections 10.5–10.6

1 Using *unrelated* samples eliminates order effects and the need to duplicate materials.

2 It introduces individual differences and requires larger numbers of participants than the equivalent related samples design.

3 We can attempt to control for individual differences by assigning our participants to conditions randomly.

4 This does not eliminate the variation introduced by individual differences; it is simply an attempt to render it *unsystematic*.

5 Assigning participants randomly to conditions is no guarantee that systematic differences between conditions will be eliminated. You should bear this in mind when interpreting your findings (and check the randomization if you know how – see the Web site accompanying this book for more on how to do this).

10.7 Matching participants

If you have the time and the resources to collect the relevant information, then **matching** your participants can give you a good halfway

house between the unrelated and related samples designs. Matching involves finding groups or even pairs of participants who are similar on some variable that you think is related to the IV. With the music and driving experiment, for instance, we might assess each participant's driving ability *before* assigning them to experimental conditions and then attempt to make sure that equal numbers of good and poor drivers are allocated to each group. Similarly, we might run the reasoning experiment by allocating equal numbers of good, poor, and indifferent reasoners to the concrete and abstract conditions. This way we do not trust to a random sequence to spread individual differences equally between conditions – instead we ensure that this happens.

Where possible, therefore, it is a good idea to match the participants in an experiment in which you have different participants in conditions, as this will reduce the possibility of individual differences confounding the effects of your independent variable. However, you will of course still have to assign matched participants to the experimental conditions *randomly*. That is, which particular condition any one of the better drivers or reasoners or any one of the less able drivers or reasoners is allocated to must be decided at random in exactly the same way as you would do with a sample of unmatched participants.

? SAQ 36 Why?

Where you have *matched* participants, therefore, you will still need to assign them to experimental conditions randomly. You do this by randomly assigning the members of each group of matched participants separately. So, for instance, you would assign the *better* drivers to their conditions using the sort of methods described in Appendix 2. Separately, you would do the equivalent with the *less* able drivers. This way you would end up with the same spread of ability in each condition but within any level of ability participants would have been assigned randomly to their conditions.

In effect, in doing this, you will have introduced an additional IV into your experiment – an unrelated samples IV comprising the better and poorer drivers. This is not a problem (in fact, usually quite the reverse): I will talk in Chapter 13 about such designs.

When you have matched *pairs* of participants, it is sometimes argued that the data can be analysed statistically as if it were from a

related sample. As a student, however, it is extremely unlikely that in human work you will achieve the necessary level of matching for this, so usually you should consider data from matched pairs of participants to be *unrelated* when analysing them.

Summary of Section 10.7

1 A good halfway house between the related and the unrelated samples design is to use unrelated samples in which the participants have been matched and assigned to conditions so that there is an equal spread of ability on the experimental task between the conditions.

2 Such designs reduce the possibility of individual differences confounding the effects of the independent variable.

10.8 External validity

With this knowledge about unrelated and related samples IVs and about matching participants you are already well on the way towards being able to design interesting experiments in psychology. In theory, there is no limit either to the number of levels that you can have on an IV or to the number of IVs that you can have in an experiment. In practice, however, there are important issues that you need to consider when using IVs with many levels or more than one IV. I will introduce you to these in the final chapter, once you have learnt about significance testing and have been introduced to the important issue of the power of your experiments. Next, however, I want us to turn to two other issues that you need to be aware of when experimenting: validity (the subject of this section and the next) and ethics (the subject of Section 10.10).

In the previous chapter we considered how the capacity of the experiment to allow causal inferences stems from the *control* that is achieved. Control involves holding other factors constant across the conditions so that they cannot account for any changes observed in the DV. Remember, we can make unequivocal causal inferences – that the IV causes the changes we see in the DV – only if we have held everything else constant.

Of course, this is a condition that can never be realistically achieved. There are just too many variables in the world. In practice, we aspire

to holding constant those variables that plausibly may be rival explanations for the observed changes in the effect variable and others that can be held constant without too much effort or too much distortion of the situation. Others we allow to vary in the hope that they will not turn out to differ systematically across the conditions. Yet others – the vast majority – we assume are irrelevant (perhaps wrongly) and disregard. Often we take it one step at a time, controlling for different variables across a series of separate experimental studies.

There is, however, an inevitable tension between experimental control and **generalizability**. Generalizability is the extent to which the findings generated by your experiment can be extrapolated. For example, are they relevant to everybody, regardless of race, class, gender, age? Or are they circumscribed in some way – for example, relevant only to middle-class, white, youngsters of above average intelligence?

The worst state of affairs is that the findings cannot be generalized to any people or circumstances beyond those used in the particular experiment that you ran. This is very unusual. However, it can be the case that the findings are of limited applicability because of some of the variables that you had to hold constant for control purposes. An obvious example would be if you choose to use only male participants (i.e., hold the variable *gender* constant across conditions); then there would be question marks over the extent to which the findings could be generalized to females.

The generalizability of the findings is also called the **external validity** of your study: *external* validity because it concerns the relevance of your findings to situations beyond or external to those used in the study. Experiments high in external validity have highly generalizable findings – that is, the findings apply to a range of people, times and situations not directly assessed in the experiment.

One threat to external validity comes from the conditions that you imposed on the time, setting and task for the purposes of control.

? SAQ 37 For instance, for some years I helped to run a class experiment at a summer school to test whether people can taste the difference between mineral water and tap water. We found that most people cannot. In this experiment, we controlled the variable *temperature* by keeping the water at room temperature. Does this mean that we can conclude that people in general cannot taste the difference between mineral water and tap water?

Another potential limit to external validity stems from your participants. Are they representative of the group from which you took them (e.g., students)? Are they representative of a wider range of people? Many studies in psychology, for example, use undergraduate psychology students as participants. Yet students, of course, tend to come from a restricted sector of the population. They are almost invariably young, supposedly intelligent, and predominantly middle-class. Add to this the probability that those interested in psychology may also have peculiarities of their own, and it could be that the findings of studies based on samples of undergraduate psychologists cannot be generalized to other populations.

However, the validity of this criticism depends on the variables in question. If the variables that *differentiate* such participants from the general public (e.g., age, intelligence, class) are thought to have some impact on the DV (e.g., reasoning, attitudes), then the criticism may be valid. However, other variables may not be so influenced (e.g., motor skills, visual perception). So, do not *automatically* assume that having a student sample seriously limits the extent to which you can generalize your findings. Always bear in mind that – if your study is in other respects well designed – the results should be generalizable to at least some subgroup in the population. After all, in some respects at least, students are still members of the human race and, of course, there are young, middle-class, intelligent people who are not students.

At first, many of you are concerned that you should always have a random sample of people from the general population in your experiments and doubt the usefulness of experiments lacking such samples. For now, remember that this is a concern with **random selection** and it relates to the *external* validity of the experiment. We will consider this issue further in a later section. But next I want to introduce you to *internal* validity.

Summary of Section 10.8

1 There is often a tension between the demands of experimental control and the extent to which you can generalize your findings to people, times, and situations not directly assessed in the experiment.

2 Experiments with highly generalizable findings are high in *external* validity.

3 Threats to external validity can come from constraints on the time, setting, task, and the participants used.

4 Concerns about the failure to randomly select a sample from a population relate to the external validity of the experiment.

10.9 Internal validity

For practical reasons, most of the experiments that you run as a student are unlikely to be very high in external validity. They must, however, be high in *internal* validity. **Internal validity** refers to the extent to which we can relate changes in the DV to the manipulation of the IV. That is, a well-designed experiment with no confounding variables is an experiment very high in internal validity. This is an experiment in which we can make the unequivocal causal inferences we desire.

However, to achieve high internal validity, we often need to hold many extraneous variables constant, thus compromising external validity. For example, in our cheese and nightmare experiment, the more we introduce controls for variables that might also produce nightmares, the more we run the risk that the generalizability of our findings will be limited by the values at which we hold those variables constant. For example, if we require people to eat cheese 3 hours before they go to bed, the findings may not apply to cheese eaten 10 minutes before going to bed. Likewise, if we restrict ourselves to one type of cheese (say, Cheddar), the findings may be limited to that type of cheese, or only to hard cheeses, or only to cheeses made with milk from British cattle, and so on.

Internal validity represents the degree of control that we exercise in the experiment. One of the key elements in achieving high internal validity (but by no means the sole one) is that we randomly assign participants to conditions (or orders of conditions to participants) in order to maximize our chances of rendering the variation on extraneous variables unsystematic between conditions. This is the issue of **random assignment** and it relates to the internal validity of the experiment.

Thus, random assignment of participants to conditions should not be confused with *random sampling* of participants from a population. At the start of your careers in psychology most of you tend to be much more concerned about the latter than you are about the former,

whereas at this stage the reverse should be the case! This is because, although both of the types of validity described above are important, internal validity is the more important. It is of little consolation to be able to generalize a set of equivocal findings to a lot of people and many situations, when you run an experiment high in external but low in internal validity. In contrast, an experiment high in internal validity usually has at least a modicum of generalizability. Of course, ideally in our studies we should aim for both. However, as students, most of the time you should concentrate your time and energies on ensuring that your experiments are at least high in internal validity. You now know how to do this: by selecting carefully the variables to hold constant across conditions, standardizing your procedures and materials, and using *truly* random procedures to attempt to ensure that other important sources of variation do not differ systematically between the conditions. Random assignment of participants to conditions (or orders of conditions to participants) is a key element in striving to achieve this.

Summary of Section 10.9

1 The experiments that you run as a student are unlikely to be very high in external validity.

2 They must, however, be high in *internal* validity.

3 In an experiment high in internal validity, we can relate changes in the DV unequivocally to the manipulation of the IV.

4 One of the key elements in achieving high internal validity is to randomly assign participants to conditions or orders of conditions to participants.

5 You should not confuse random assignment of participants to conditions with random sampling of participants from a population. The former relates to the internal validity of the experiment, the latter to the external validity of the experiment.

10.10 Ethics: The self-esteem and well-being of your participants

A concern with variables and control is not the only element of experimental design in psychology. Regard for the welfare, rights,

and dignity of your participants is as essential to good experimental practice as is the pursuit of experimental rigour and elegance. Never underestimate the impact that being in a psychological experiment can have on your participant. We live in a competitive society and the experiment has all the trappings of a test. Those of you who have participated in a psychological experiment will probably have been struck by just how difficult it can be to shake off the feeling that somehow you are under scrutiny. You can feel that you have to do your best, even though you know that an experimenter's interest is rarely in particular individuals, but invariably in making statements about the influence of the IV on *groups* of people. Those of you who have not participated in one should try to if possible. It gives you an invaluable insight into the sorts of factors that can influence a participant's performance in an experiment.

When designing any study it is important to have clearly in mind the participants' welfare, self-respect and dignity. You must not put people through an experience that undermines their welfare or self-regard. So, think about the impact that your study will have on your participant when you think about designs that you might employ, measures that you might use, and issues that you might address. For example, look closely at your procedures, the wording of your instructions, the items on any questionnaires or other measures that you are thinking about using. Is it really appropriate and sensitive to use the depression inventory that you are considering? Could some of the questions on the anxiety scale that you are thinking of using cause embarrassment or distress? If so, are there better alternatives? If not, is your use of these measures really justified and ethically acceptable? (While we are on the topic of published measures, such as scales and tests, make sure that you have permission to copy and use any published measures that you want to use in your study, including any you have downloaded from the Internet. See your tutor for advice on this.)

Even as students, your conduct is expected to adhere to ethical principles. In your early work in practical classes the tutors running the course will have primary responsibility for ensuring that the things they ask you to do are ethically acceptable. However, when you undertake your own project work (e.g., in the UK in the final year of an undergraduate degree), the likelihood is that you will be expected to submit details of your proposed study to an ethics committee for prior assessment. You will be allowed to run your study only once you have received approval from that ethics committee.

It is important, therefore, that you get hold of a copy of the ethical code that is used in your department or school. You must familiarize yourself with these from the outset. When the time comes, you should also make sure that you know when and how to submit your studies to the relevant ethics committee and that you do not undertake work that needs prior ethical approval without having that approval. If in any doubt, check with your tutor.

In the unlikely event that your department or school does not have a set of **ethical principles** then you can obtain some from the professional body that oversees the conduct of properly qualified psychologists in your country. These are organizations such as the British Psychological Society, the Australian Psychological Society and the American Psychological Association. You can usually download these principles from the organization's Web site. Failing that, you can ask them to send you a copy.

The principles deal with a number of aspects of conducting yourself ethically with participants. These include:

- approaching people and encouraging them to take part in your study;
- what you do to them and how you interact with them during the study;
- how you respond if they decline to take part or wish to withdraw once they have started;
- what you say to them after they have finished;
- how you safeguard and report the data that you obtain from them.

You must make sure that you have read the most recent and relevant set of principles and that you abide by them.

10.10.1 Informed consent

A key feature of all ethical principles is that of **informed consent**. Informed consent is what it sounds like – telling people enough about your study to enable them to make an *informed* decision about whether or not to take part. For example, the *Publication Manual of the American Psychological Association* (APA, 2001, p. 391) states that psychologists should "use language that is reasonably understandable to research participants" to do the following:

- inform them of the nature of the research;
- inform them that they are free to participate or to decline to participate or to withdraw from the research;
- explain the foreseeable consequences of declining or withdrawing;
- inform participants of significant factors that may be expected to influence their willingness to participate (such as risks, discomfort, adverse effects);
- explain other aspects about which the prospective participants inquire.

Standard 8.02 of the APA's new ethics code (APA, 2002) is similar, as is principle 3.3 of the British Psychological Society's *Ethical Principles* (BPS, 2006).

The principle of informed consent is very important. However, respond to it sensibly. It is not a directive to spell out *everything* about the study to prospective participants. Given that people would respond very differently in many studies if they knew *precisely* what it was about, you must think carefully about what you need to tell them about your study. It is important to give your participants a general idea about what you will ask them to do, how long it will take (honestly!), and so on. It is *essential* that you tell them about risks, discomfort, or any other aspects that might affect their willingness to take part. (Indeed, the ethics committee will insist on this as a condition of approval.) However, think carefully about what you tell potential recruits about what the study is *about*. Tell them the things that they need to know in order to make an informed decision. Do not jeopardize the study unnecessarily by telling them things that are important to interpreting the experiment and yet *not* key to whether or not they will agree to take part. (As ever, take your tutor's advice on this.)

Do not confuse *masking* or *concealing* or *withholding* the hypothesis with *deception*. **Withholding the hypothesis** involves encouraging participants to suspend judgement about what you are investigating until after the study has finished and you are ready to debrief them. **Deception** involves deliberately misleading participants into believing that the study is about something it is not about. Participants generally are quite willing to do the former and can be pretty aggrieved by the latter. As a student, you must not run a study involving deception without prior ethical approval and *extremely* close supervision by a tutor.

10.10.2 Debriefing your participants

You must make time available for those who want it at the end of your study to discuss any questions that they may have regarding what you have asked them to do. This is known in the trade as **debriefing** your participant, and it is an integral part of good research practice. At this stage, answer their questions informatively, honestly, and willingly, discussing the thinking behind your ideas, and the purpose of the study if that interests them. If you cannot tell them precisely what the study is about, for fear that word will get about and future participants will no longer be naive about the study, then see if they are prepared to wait for full debriefing until all the participants have taken part. If so, you should take contact details and send them an account of the study and the findings as soon as these are available. Debriefing is for the *participant's* benefit, not yours.

10.10.3 Studies on the Internet

You may be tempted to run a study on the Internet. If so, then you will need to work very closely with your tutor to ensure that the study is ethical. This is because the Internet presents a number of significant problems for the principles we considered above. For example, it can be difficult to get genuinely informed consent and to debrief all participants properly, especially those who quit your study before the end. Because you do not have face-to-face contact with the participant, you cannot monitor, support or terminate the study if it upsets them. Likewise, you have no control over the conditions under which they view and interact with your stimulus materials, and – as you now appreciate – rigorous control is key to good experimenting. For these and other reasons, you should think very carefully before deciding to run your study on the Internet, make sure that you familiarize yourself with ethical guidelines for running Internet studies before you do, and make sure that you work closely with a tutor at all stages.

10.10.4 Data confidentiality

Ethical considerations do not finish with the completion of the study. Most countries now have regulations regarding how data on human beings are stored and professional guidelines also cover this aspect of

conduct. You must keep the data safe, secure, and confidential, no matter how dull, boring and uninteresting they may seem to you to be. This includes both the original measures (such as a set of questionnaires or response sheets) and the data that you have on computer. Consult the relevant guidelines to learn more.

Summary of Section 10.10

1 Regard for the welfare and comfort of the participants in your study is an integral part of good design. You must not put your participants through an experience that undermines their welfare, self-respect or dignity, and you must bear this in mind when thinking about procedures that you might employ and issues that you might address in your studies.

2 You should be aware of the contents of the most relevant set of ethical principles before you start designing and running your own studies. Even as students, your conduct must adhere to ethical principles.

3 If you go on to do project work, you may need to submit your proposed study to an ethics committee for approval. If so, you must not start the study before receiving ethical approval.

4 A key feature of all ethical principles for research with human beings is the principle of informed consent. Make sure that you understand this principle and abide by it.

5 Make time available at the end of your experiments to discuss any questions that your participants have regarding what you asked them to do. You should answer their questions honestly, openly, and willingly. This is known in the trade as debriefing your participant, and it is an essential feature of any well-run experiment.

6 Exercise care if you wish to run a study on the Internet, as Internet studies raise a number of significant ethical issues.

7 Keep all data you collect safe, secure, and confidential and in accordance with the laws governing the storage of information about humans.

Consolidating your learning

This closing section is designed to help you check your understanding and determine what you might need to study further. You can use the list of methodological and statistical concepts to check which ones you are happy with and which you may need to double-check. You can use this list to find the relevant terms in the chapter and also in other textbooks of design or statistics. In this chapter we also discussed ethical issues and below you will find key ethical terms. The diagnostic questions provide a further test of your understanding. Finally, there is a description of what is covered in the two statistics textbooks paired with this book.

Methodological and statistical concepts covered in this chapter

Carry-over effect
Counterbalancing
External validity
Extraneous variation
Generalizability
Individual differences
Internal validity
Matching participants
Order effect
Practice effect
Random assignment
Random selection
Randomizing
Related samples (within subjects, within participants, dependent/correlated samples, dependent/correlated/repeated measures) IV
Unrelated samples (between subjects/groups, between participants, independent/uncorrelated samples, independent/uncorrelated measures) IV

Ethical concepts covered in this chapter

Data confidentiality
Debriefing
Deception
Informed consent
Withholding the hypothesis

Diagnostic questions – check your understanding

1 What is a *related* samples IV?
2 What is an *unrelated* samples IV?
3 What is *random assignment* to conditions?
4 What is the *internal validity* of an experiment?
5 What is the *external validity* of an experiment?

You can find the answers at the rear of the book (p. 276).

Statistics textbooks

∑ The books paired with *Designing and Reporting Experiments* have the following coverage:

Learning to use statistical tests in psychology – Greene and D'Oliveira
Greene and D'Oliveira discuss issues involved in using the same or different participants in the conditions of the experiment in Chapter 2.

SPSS survival manual – Pallant
Note that Pallant uses "independent samples" or "between groups" to describe unrelated samples IVs and "paired samples", "repeated measures" and "within subjects" to describe related samples IVs. She describes in Part Five a wide range of statistical methods suitable for analyzing data from studies like those I have described in this chapter; however, you should not consider using any of these until you are familiar with the material in Chapter 11.

11 Statistics: significance testing

So far in this book I have talked rather glibly about *differences* between conditions and *correlations* between variables. However, what I have really meant here is what you will come to understand as differences and correlations that are **reliable**. For there is something special about the differences and correlations that interest us. We are not interested simply in, say, a difference in the numbers between two conditions; we are interested in those differences (and correlations) that would usually be repeated if we were to run the study again.

Thinking back to our driving and music experiment in Chapter 10, if we were to measure the performance of any given participant on the driving simulator on a number of occasions, we would not expect him or her to obtain an *identical* number of errors on each occasion. We would not even expect this when s/he was fully accustomed to the apparatus. Similarly, I have already pointed out that one of the problems with *unrelated* samples is that there are generally differences between people in their basic abilities on experimental tasks. This means, of course, that even when we assign participants to experimental conditions randomly, we would hardly expect our different groups to obtain identical mean values on the dependent variable.

If you have any doubts about this, then try the following exercise. Try balancing a book on your head. (Make sure that it is not too heavy, and keep away from breakable objects!) See how long you can keep it there (a) in silence and (b) with the radio on. Do it five times under each condition. (Make sure that you set up a *random* sequence *in advance*.) Record your times, and then work out the *mean* time for balancing under both conditions. Are they identical? I bet they are

not! Does this mean that having the radio on has affected your ability to balance a book on your head? (I trust that you have controlled adequately for the effects of practice and fatigue.)

The point of this is that you will more or less inevitably find differences between conditions on the DV in an experiment, *even when the IV does not in fact cause changes in the DV*. So, for example, even if listening to music has absolutely no impact whatsoever on the ability to drive, we would still expect to find our participants driving at least slightly better in one of the conditions of our experiment than another. That is, we will always have to assess the effects of our independent variable against a background of *inherent* variation.

We need, therefore, to find a way of being able to discriminate a difference that *has* been influenced by our IV (a *reliable* difference) from one that has *not* been so influenced – a difference that would be there anyway. That is, we need to be able to detect the sort of difference that would occur simply if we took a group of participants and measured their performances on repeated occasions *without* exposing them to the IV. This is why we turn to *inferential statistics*.

Summary

1 We will invariably find that there are differences on the DV between the conditions even when the IV does not cause changes in the DV.

2 We therefore have to find some way of distinguishing a difference that has been caused by the IV (a *reliable* difference) from one that is simply the product of chance variation.

3 We employ *inferential* statistics to assist us with this task.

11.1 Inferential statistics

Think back to our cheese and nightmare experiment (Chapter 9). Imagine that we actually ran this experiment one night and obtained the data in Table 11.1. (There are 50 participants in each condition.) You can see from Table 11.1 that 11 more of those in the cheese condition reported nightmares than did those in the no cheese condition. Thus, *more* people reported nightmares in the cheese condition than in the no cheese condition. This is a difference, but as yet we do

Table 11.1
The Number of Participants Reporting Nightmares and not Reporting Nightmares in the Cheese and No Cheese Conditions

	Nightmare	
Condition	Yes	No
Cheese	33	17
No cheese	22	28

not know whether it is the sort of difference that we would be likely to find anyway. The question is, if eating cheese actually has *no effect* upon the incidence of nightmares, how likely would we be to obtain a difference of 11 between our conditions?

To understand this, imagine that we took 100 table tennis balls and stamped "nightmare" on half of them and "no nightmare" on the remainder. We then placed these in a large black plastic bag, shook them up, and drew them out one by one. Imagine also that we simply decided to assign the first 50 table tennis balls we drew out to a hypothetical *cheese* condition and the remainder to a hypothetical *no cheese* condition. The question is, doing this, how likely would we be to get the sorts of values that we have in Table 11.1?

Ideally we need some way of assessing this. This is precisely where **inferential statistics** come in. These statistics enable us to assess the likelihood that we would obtain results like ours *if* we assume that our IV did *not* cause changes in our DV. That is, using such statistics we can assess whether we would have been likely to get data like ours *if* we assume that there is not a genuine effect of eating cheese on the incidence of nightmares.

You may have spotted that the assumption that the IV does *not* cause changes in the DV is one that we have met before. It is, of course, the *null hypothesis* (Section 9.1.4). Essentially, therefore, inferential statistics tell us the probability that we would obtain results like ours *if we assume that the null hypothesis is true*. This is what we call the probability *under the null hypothesis*.

There are inferential statistical tests available to assess this probability for all kinds of data. For all of these tests, however, at the end of the calculations you should end up with a single score. This score is the value of the statistic that "belongs" to your data. Associated with

Table 11.2
Probabilities Under the Null Hypothesis Associated with Particular Values of Chi-Square with One Degree of Freedom

Chi-square	Probability
0.016	.90
0.15	.70
0.46	.50
1.64	.20
3.84	.05
5.02	.025
10.83	.001

this statistic is a *probability*. This is the probability we are interested in. It tells us how likely we would be to obtain our value – and, by implication, these data – if we assume the null hypothesis to be true. Statistical software packages usually display this probability for you, along with the statistic. If you have calculated the statistic by hand, however, you can look up the probability associated with your calculated statistic in the appropriate tables of *critical values* (to be found at the rear of most textbooks of statistics – see Appendix 3).

For instance, as we have only one observation from each participant (nightmare or no nightmare), an appropriate statistic for the above example is chi-square (Section 4.6.1). Chi-square tests for an *association* between two variables – in this case whether there is any association between eating cheese and the numbers of nightmares reported. So, you can see that it is eminently suited for the question that we wish to ask of our data.

Table 11.2 contains a range of values of chi-square, together with their associated probability if we assume the null hypothesis to be true. You can see that, under this assumption, a value of $\chi^2 = 0.15$ has a probability of occurring of $p = .7$, which is quite high. A value of $\chi^2 = 3.84$ however, has an associated probability of only $p = .05$, which is quite low.

For those new to the idea of **probability**, it is measured conventionally from one to zero. An event with a probability of one is *inevitable*. An event with a probability of zero is *impossible*. Most events in our universe tend to lie somewhere between these two extremes.

If you enter the data in Table 11.1 into your statistics package it will calculate for you the value of chi-square that belongs to these

data. Alternatively, you can find details of how to calculate chi-square by hand in textbooks of statistics (see, for example, Greene & D'Oliveira, Chapter 20). Either way, you should obtain a value of $\chi^2 = 4.89$. What you need to establish next is whether this value is *likely* or *unlikely* to have occurred under the null hypothesis – that is, even if cheese does not cause nightmares.

If you calculated chi-square using a statistical software package, the package will give you the *exact* probability associated with the statistic. This value is $p = .03$. Those of you who worked out the value of chi-square by hand can establish from Table 11.2 that the value of $\chi^2 = 4.89$, with 1 degree of freedom, has a probability under the null hypothesis of somewhere between $p = .05$ and $p = .025$. For, the value of chi-square that has a probability of $p = .05$ is 3.84 which is *less* than our obtained value of chi-square. However, the value of chi-square that has a probability of $p = .025$ is 5.02 which is *greater* than our obtained value. So, our value lies somewhere in between these and consequently the probability associated with it must lie somewhere between $p = .05$ and $p = .025$. This agrees with the $p = .03$ that we obtained from our statistics package but is less exact. (See Appendix 3 for more on how to work out the probability associated with your statistic when you do not use a statistics package.)

The probability of obtaining a value of chi-square as *large* as 4.89 under the null hypothesis, therefore, is 3 times in 100 (this is just another way of expressing $p = .03$). Essentially, therefore, if we were to draw our table tennis balls randomly from our big black plastic bag we would expect, *on average*, to obtain data like ours 3 times for every 100 times that we did it.

? SAQ 38 Our obtained value of chi-square has an associated probability of $p = .03$. Is this nearer the inevitable or the impossible end of the continuum? (That is, is it nearer one or zero?)

Let's recap. We've run our experiment and generated the data in Table 11.1. We then realized that in order to make any sense of it – in order to find out what it tells us about the relationship between eating cheese and experiencing nightmares – we needed to discover how likely we would be to have obtained results like these anyway, even if

eating cheese did not cause nightmares. So, we employed a statistical test appropriate for our data (chi-square) and we discovered that the value of chi-square that we obtained (4.89) has a low probability of having occurred *if* we assume that eating cheese *does not* induce nightmares. Does this mean, therefore, that we can safely conclude that eating cheese *does* induce nightmares?

Well, the simple and rather unsettling answer to that question is – no. For the problem is that we still do not know whether the IV does or does not affect the DV. All we know is that the probability of getting data like ours is relatively low *if* we assume that the IV does *not* affect the DV. Thus, under the null hypothesis, data like ours are *unlikely*. However, they are still *possible*. We know, therefore, that on any given occasion we would be unlikely to draw a similar distribution of values to those in Table 11.1 from our big black plastic bag of table tennis balls. Nevertheless, we might actually do so. We have, in fact, about a 33-1 chance of doing so. Moreover, as some of us know to our (very!) occasional benefit, long shots, though infrequently, still win races.

In fact, we will *never* know whether the null hypothesis is true or false. Inferential statistics do not provide us with sudden insights into the laws of the universe. They simply tell us the probability that we would get values like those we obtained on our DV *if* the null hypothesis *were* true.

So how do we get around this problem? Well, all is not lost. The procedure that we traditionally adopt to resolve this dilemma is described next.

11.2 Testing for statistical significance

In this procedure, what we do is set up a criterion probability for our statistic. We do this before we even begin collecting the data. We decide subsequently to obey a simple rule. The rule is as follows. If we find that the *probability* associated with our obtained statistic falls *at* or *below* this criterion (i.e., is the same value as the criterion probability, or nearer zero), then we will *reject* the assumption that the null hypothesis is true (see Figure 11.1). If, however, we find that the probability associated with the statistic falls *above* this criterion (i.e., is nearer one) then we will *not* reject the assumption that the null hypothesis is true.

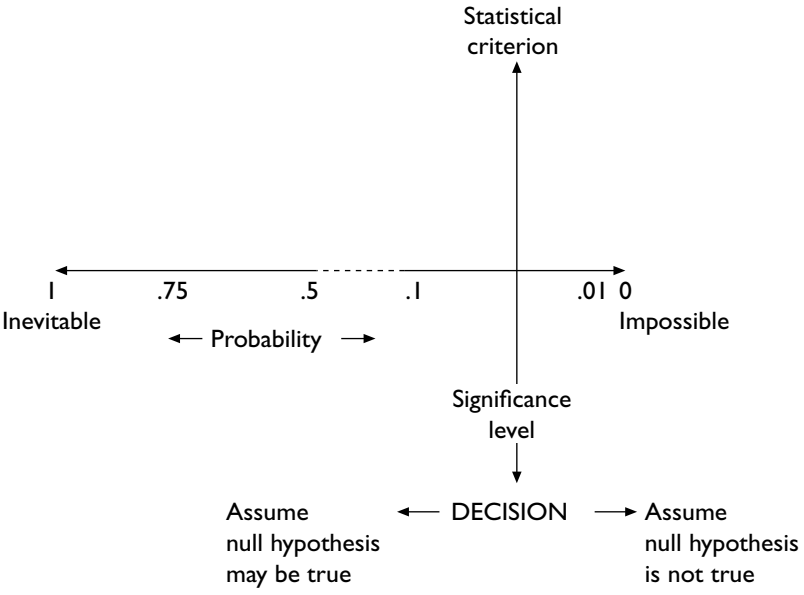

Figure 11.1. Statistical inference. If the probability associated with the obtained statistic falls to the left of the statistical criterion (i.e., nearer $p = 1$), then assume that the null hypothesis may be true. If it is the same as the criterion probability or falls to the right (i.e., nearer $p = 0$), then assume that the null hypothesis is not true.

This is the principle of **statistical significance** testing. Using this approach, we *always* do this when we test our data for differences between conditions. The criterion is referred to as the **significance level** or **alpha level**. Although we are free to vary it, this level is often set at a probability of less than or equal to .05 (in other words 5 times in a 100 or 1 in 20). This is known as the **5 percent significance level** (Figure 11.2).

What we do, therefore, is compare the probability associated with the obtained value of the statistic with our chosen significance level. If this probability *equals* or is *less* than the significance level (i.e., is nearer zero, a less probable event), then we decide that we have sufficient evidence to **reject the null hypothesis**. Such data we describe as **statistically significant**. Under these circumstances we would talk of a "statistically significant difference" between our conditions. If, however, the obtained statistic has a probability *greater* than our

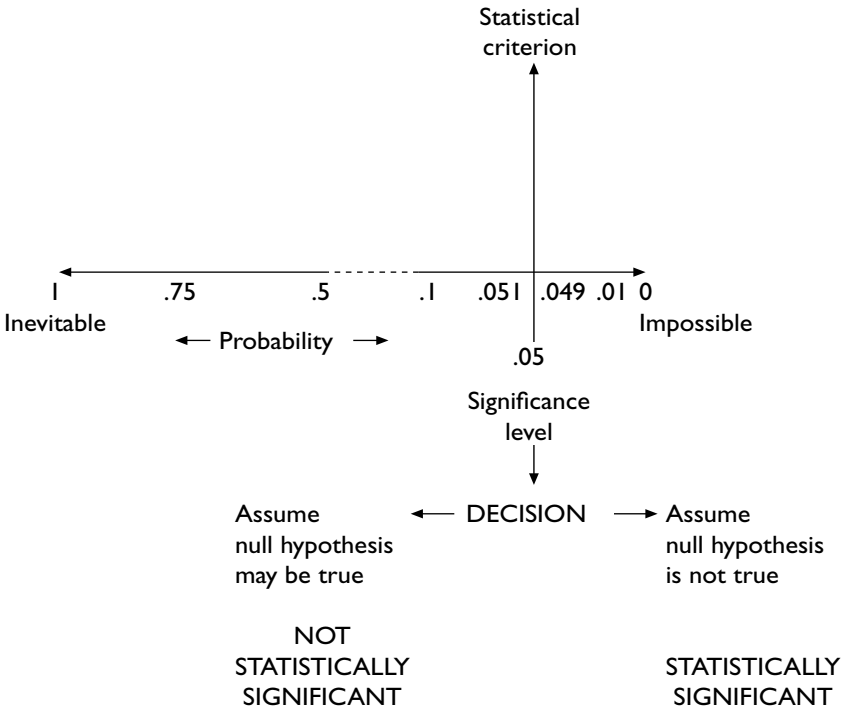

Figure 11.2. The 5 percent significance level. If the probability associated with the obtained statistic is greater than $p = .05$, do not reject the null hypothesis. The data are not statistically significant. If this probability is equal to or less than $p = .05$, then reject the null hypothesis. The data are statistically significant.

significance level (i.e., is nearer one, a more probable event) then we decide that we have insufficient evidence to reject the null hypothesis. We thus **fail to reject the null hypothesis**. Findings such as these we describe as **not statistically significant**.

This terminology, though widely used, is nevertheless unfortunate. To call our results *statistically* significant does not necessarily mean that they have much *psychological* importance. Likewise, results that are not statistically significant may yet be psychologically informative. Significance, in this context, is a statistical concept. It tells us something about the statistical nature of our data. The psychological importance of a set of findings, whether statistically significant or not, remains to be established. When working with inferential statistics it

is important, therefore, to differentiate in your mind the concepts of statistical significance and psychological importance. Do not assume that the former implies the latter. I have more to say about this in Chapter 12.

In the strict application of this procedure, once you have set up the significance level, you are compelled to abide by the outcome. So, if the probability associated with your obtained statistic turns out to be *greater* than the significance level (even if only by a minute amount), then you must assume there is not enough evidence to reject the null hypothesis. I will have more to say about this in the next section.

In the case of our cheese and nightmare experiment, we calculated a value of $\chi^2 = 4.89$. We now need to find out whether this is statistically significant at the 5 percent significance level. (This is the significance level that I chose on our behalf before collecting the data.) This means that we need to work out whether the probability associated with $\chi^2 = 4.89$ is less than or equal to $p = .05$.

? SAQ 39 Earlier we established that the probability associated with $\chi^2 = 4.89$ with 1 degree of freedom is $p = .03$. Is our value of χ^2, therefore, statistically significant at the 5 percent significance level?

In effect, therefore, at the end of our experiment, having analysed the data, we find ourselves in a position to make a decision regarding the null hypothesis – that is, whether to reject it or not to reject it. We do this on the basis of whether or not the probability associated with our obtained statistic is less than or equal to the significance level we set prior to running the experiment.

Summary of Section 11.1–11.2

1 *Inferential* statistics tell us how likely we would be to obtain data like ours *if* the IV did *not* cause changes in the DV. That is, they tell us this probability *assuming the null hypothesis is true*. Such statistics do not, however, tell us whether or not the null hypothesis really *is* true.

2 In fact, we can never know for certain whether the null hypothesis is true or not true.

3 One way around this problem is to adopt a criterion probability. We call this the *significance level* or *alpha level.*

4 Although we are at liberty to alter the significance level, it is set conventionally at $p = .05$. This is known as the "5 percent significance level".

5 If the *probability* associated with our obtained statistic is *less than* or *equal to* our significance level, we decide to reject the assumption that the null hypothesis is true. Such results we describe as *statistically significant*.

6 If the probability associated with our obtained statistic is *greater than* our significance level, we decide *not to reject* the assumption that the null hypothesis is true. Such results we describe as *not statistically significant*.

7 In this context, significance is a statistical concept. To say that a difference is *statistically* significant does not mean that *psychologically* the difference has much theoretical or practical importance. Likewise, the absence of a statistically significant difference may nevertheless be of theoretical or practical importance.

11.3 Type I and type II errors

At the end of our experiment, therefore, we will have decided whether or not to reject the null hypothesis. However, the simple fact is – *we could always be wrong*. That is, we may find ourselves rejecting the null hypothesis when – had we seen the tablets of stone on which are written the Laws of the Universe – we would find that the null hypothesis should *not* have been rejected. For instance, in the case of our cheese and nightmare experiment, our obtained value of chi-square is 4.89, which has an associated probability that is less than the conventional significance level of $p = .05$. Using this significance level, therefore, we would have to reject the null hypothesis. However, suppose that you receive a visitation from the powers that run the universe, who reveal to you that actually the null hypothesis is in this case correct. Under these circumstances we would have *rejected* the null hypothesis when we should not have done. This is known as a **type I error**. Moreover, not only is it an integral feature of the process of statistical inference – of acting as if we know something for

certain when we do not – but, in this world at least, we never know whether we have made it (visitations from the powers that run the universe aside). However, we do know the *probability* that we have.

What is the probability of making a type I error (i.e., of incorrectly rejecting the null hypothesis)? The probability of making a type I error (if the null hypothesis is true) is, in fact, the significance level. We decide to reject the null hypothesis when the probability associated with our obtained statistic reaches the significance level. Under these circumstances we *always* reject the null hypothesis. On some of these occasions, however, unbeknown to us the null hypothesis will be true. With the 5 percent significance level, therefore, because we reject the null hypothesis *every* time that our data reaches this level, we will make a mistake, on average, once in every 20 times that we do it.

So, what can we do about this? After all, we would like to minimize our mistakes. We do not want *too* many of the factual assertions that we make about the psychological universe to be wrong. Well, if our significance level is a measure of our type I error rate, then one thing that we might do is to reduce the probability of making this error by *making our significance level more stringent* (Figure 11.3). We make our significance level more stringent by choosing to use a *lower* probability under the null hypothesis to be the significance level – such as $p = .01$. A significance level of $p = .01$ has a type I error rate of only 1 in 100, rather than the 5 in 100 (or 1 in 20) of the $p = .05$ significance level.

Now, doing this is not necessarily wrong. However, although we can get the probability of making a type I error *closer* to zero, the only way that we can actually *make* it zero is by saying nothing at all about the existence or non-existence of a reliable difference between the conditions in our experiment. If we wish to say anything at all, therefore, we have to accept some probability of making a type I error.

Making the significance level more stringent is a cautious or, as we say in the trade, a *conservative* thing to do. Sometimes it may be the appropriate step to take. For example, those who undertake research into extrasensory perception tend routinely to employ more stringent significance levels in order to reduce their chances of making a type I error. However, our job as scientists is essentially to make the best guesses we can about the existence of causal relationships. Making the significance level more stringent may well reduce our tendency to make mistakes, but it does so at the cost of reducing our ability to say anything. Take it too far and we are in danger of throwing the scientific baby out with the statistical bath water.

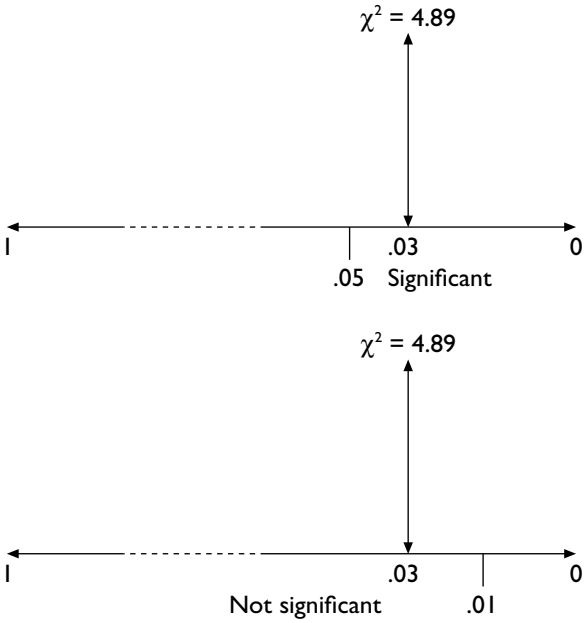

Figure 11.3. The effects of making the significance level more stringent. A $\chi^2 = 4.89$ with 1 degree of freedom is statistically significant at the 5 per cent significance level but is not statistically significant at the 1 per cent significance level.

To put it another way, in making the significance level more stringent, we are indeed reducing the probability of making a type I error. That is, we reduce our chances of thinking there is a reliable difference when there is not one. However, at the same time, we *increase* the probability of making another type of error – that is, of *not rejecting* the null hypothesis when we should have done. This is known as a **type II error**. When we make a type II error, we *fail* to detect a reliable difference that *is* there.

For instance, suppose that we had adopted the $p = .01$ level of significance for our cheese and nightmare experiment. With $\chi^2 = 4.89$ (our obtained value), we would not have achieved statistical significance, and so would have been compelled not to reject the null hypothesis (Figure 11.3). However, suppose that on the tablets of stone containing the Laws of the Universe was written that eating cheese will induce nightmares in humans. By making our criterion too

stringent, we would therefore have missed the opportunity to make a correct statement about the universe in which we live. I will have more to say about type II errors in Chapter 12.

Summary of Section 11.3

1 Sometimes we will reject the null hypothesis when it is in fact true. This is known as a *type I error*. We cannot avoid making such errors if we wish to say anything positive about the psychological universe.

2 If the null hypothesis is true, then the probability of making a *type I error* is the *significance level*.

3 We can, therefore, reduce the likelihood of making a type I error by using a more stringent significance level. However, this *increases* the probability of making a *type II error* – the error of *not* rejecting the null hypothesis when it is in fact false.

4 The conventional significance levels ($p = .05$ and $p = .01$) are compromises between the rates for these two types of error.

11.4 Choosing a statistical test

This, then, is the procedure in principle. It is the same for all tests of statistical significance. However, how you realize the procedure for testing for statistical significance in practice depends on the precise test that you need to employ to analyse your data. You will have to employ different tests of significance under different circumstances if you wish to arrive at a meaningful answer to the same question: should I reject the null hypothesis in my experiment? Even if you are not using the test for the purposes of significance testing (see Chapter 12), you still need to know the most appropriate procedure for obtaining the statistics you require.

Choosing the appropriate statistical test is the key to the whole process of finding out exactly what your data have to tell you. However, it is nowhere near as daunting a process as it may seem. (I know, we all say this, but believe me it *is* true.) If you keep a clear head and take things *step by step*, most of the time you should have few problems in arriving at an appropriate test for your data. Indeed, your problems will be reduced in this respect if you get into the habit

of thinking about how you are going to analyse your data *before* you start running your study. You can then modify the design if you foresee any problems.

In order to decide which test to employ, you need to be able to answer a number of questions about the *type of study* you conducted, the *nature of the data* you obtained, and the *precise questions that you wish to ask* of these data. Below is a list of the key questions that you need to answer:

1 Do you want to (a) compare conditions or (b) correlate variables?
2 What type of data do you have? Are they (a) nominal, (b) ordinal or (c) at least interval?
3 If at least interval, can you assume that the data come from a population of scores that is normally distributed?
4 If you wish to compare conditions, did you manipulate more than one IV in your study?
5 If you wish to compare conditions, how many do you wish to compare at any one time?
6 Do you have the *same* or *different* participants in these conditions?

You can find more about each of these questions in Section A of the book's Web site at **http://mcgraw-hill.co.uk/openup/harris/**. They are also discussed in the statistics textbooks paired with this book (see the end of this chapter).

One problem to watch out for when writing your report is that there can sometimes be more than one way of analysing the same set of data. That is, it is not uncommon to find that you can test the *same* prediction using the *same* dependent variable but with *different* statistical tests. For example, the independent *t* test and the Mann-Whitney U test are alternative tests for data from an unrelated samples IV with two conditions. For the purposes of your report, however, you should not duplicate tests. That is, under these circumstances you should use and report only one of the comparable tests. (For example, either the independent *t* test or the Mann Whitney U test.)

Summary of Section 11.4

1 Choosing the appropriate statistical test is the key to the whole process of arriving at a meaningful answer to the question: should I *reject* or *not reject* the null hypothesis in my experiment?

2 In order to decide which test to employ, you will need to address a number of questions about the type of study, the nature of the data, and the precise questions that you wish to ask.

3 Often there may be more than one way of testing statistically the same prediction using the same dependent variable. Under these circumstances, however, you should report the outcome of only one of the tests.

11.5 Two-tailed and one-tailed tests

With some statistics you need to decide whether to use a **two-tailed** or a **one-tailed test** of significance. This relates to whether or not you consider the predictions derived from your experimental hypothesis to be directional or nondirectional and it introduces us to another controversy.

Earlier I said that you should consider your predictions to be directional if you were able (for sound theoretical reasons) to specify the *direction* in which the difference between your conditions should occur. Some authors argue that this translates directly into whether or not you should employ a one- or a two-tailed test of significance.

The difference between a one- and a two-tailed test of significance has to do with the way in which you use the tails of the distributions of the statistic you are employing. What it means in practice is that you can obtain significance with data using a one-tailed test that would fail to reach significance with a two-tailed test. This is the reward for specifying the direction in which the postulated difference will occur.

Now, you may read elsewhere that whether or not you choose to employ a one-tailed or a two-tailed test depends on whether the predictions are directional or nondirectional. That is, on whether or not you have reasonable grounds for specifying in advance that one of the conditions will exceed another on the DV. Unfortunately, however, it is not quite as simple as that.

I recommend that you employ the one-tailed test *only* when you have asked a "whether or not" question of the data. That is, a question of the form, "whether or not" a particular teaching method improves the rate at which people learn a particular task, or "whether or not" a particular set of work conditions improves output. Under these circumstances, if you find a difference between the conditions *opposite*

to the one you are testing this would mean the same to you as if you had found *no difference* between the conditions. That is, if you found that the teaching method actually *impairs* the rate at which people learn the task, or the work conditions actually *decrease* output, this would be the same to you as finding that the new teaching method was no quicker than the old one, or that production was similar under the usual working conditions.

In practice, such questions generally concern those attempting to solve an *applied* problem. As students, however, most of the time you will be running studies in which findings that contradict the predictions will be every bit as important as findings that are consistent with the predictions. So, most of the time you will also be interested in findings that go in the opposite direction to the one you specified and your test should therefore be two-tailed. Only when your experimental question can simply be either confirmed or refuted – when it is open to a simple yes or no answer – should you employ the one-tailed test of significance.

If you are using tables of statistics, most contain the probability values for two-tailed comparisons. For many of the tests that you are likely to use as a student, you can derive the one-tailed value simply by halving the significance level. So, if the value that you obtain is significant at the 5 percent level (two-tailed), then the significance for a one-tailed test is actually 2.5 percent. Check in your textbook of statistics whether this is the case for the statistic that you are using.

? SAQ 40 The value of $t = 1.73$ is the critical value of t with 20 degrees of freedom at the 5 percent significance level for a one-tailed test. What is its associated probability for a two-tailed test?

One final thing to bear in mind about your analyses is that statistics only deal with the numbers fed into them. A statistical package will churn around any set of numbers that you feed into it in a suitable format. Just because the data were analysed does not mean either that the analysis itself or its outcome was necessarily meaningful. For instance, the value of chi-square for the cheese and nightmare experiment is perfectly reasonable, given the numbers that I fed into the analysis. Nevertheless, this does not tell us *anything* about the relationship between eating cheese and experiencing nightmares – I made

the numbers up! Bear in mind also that rejecting the null hypothesis does not automatically entail that the results went in the direction you predicted. I am constantly amazed how few students actually *look at* their data once they have analysed them, or think much at all about what the results really mean. Get into the habit of going back to your data and thinking about what the outcome of your analyses might mean. (This is particularly a problem when you reject the null hypothesis but fail to spot that the difference is actually *opposite* in direction to the one you predicted with a directional hypothesis. You would be surprised how often this happens.)

Summary of Section 11.5

1 Tests of significance can be one-tailed or two-tailed. The difference in practice between these versions of the same test is that you can obtain statistical significance using the one-tailed version, with data that would fail to reach statistical significance using the two-tailed version.

2 You should only use one-tailed tests when you are able to *ignore* differences *opposite* in direction to those predicted by the experimental hypothesis. When you must not ignore such differences, the test is two-tailed. Most of you, most of the time, will therefore be undertaking two-tailed tests.

3 Just because the data have been analysed statistically does not mean that the analysis or its outcome is necessarily meaningful. Get into the habit of going back to the data to interpret the outcomes of the analyses and to ensure that these outcomes appear to make sense. In particular, watch out for occasions when you have been compelled to *reject* the null hypothesis and yet the differences between the conditions are in fact *opposite* in direction to the ones that you predicted.

11.6 Testing for statistical significance: summary of the procedure

1 Decide whether the *inferential* question involves *comparing* conditions or *correlating* variables.

2 Choose a test that asks this question and that is appropriate for the type of data you will collect.
3 Before starting, set your significance level and decide whether the test will be one- or two-tailed.
4 Calculate the *obtained* value of the above statistic, whether using a statistical package or by hand.
5 With statistical packages the software will print out for you the exact probability associated with the obtained value. Check whether this probability is equal to or less than the significance level you decided to use before the experiment (commonly $p = .05$). If it is, then the data are statistically significant. If it is not, then the data are not statistically significant.
6 If calculating by hand, look in the appropriate set of tables for the statistic used, at the significance level you have adopted, to find the relevant *critical* value of the statistic (Appendix 3). If the *obtained* value of the statistic is *equal* to the critical value or differs from it in the direction required for statistical significance, then the data are statistically significant. Under these circumstances, *reject* the null hypothesis. If not, then the data are not statistically significant. Under these circumstances, *do not reject* the null hypothesis.
7 Look again at the descriptive statistics and think about why the data have compelled you to reject or not reject the null hypothesis. What, if anything, does this tell you about the relationship between the IVs and DVs?

Consolidating your learning

This closing section is designed to help you check your understanding and determine what you might need to study further. You can use the list of methodological and statistical concepts to check which ones you are happy with and which you may need to double-check. You can use this list to find the relevant terms in the chapter and also in other textbooks of design or statistics. The diagnostic questions provide a further test of your understanding. Finally, there is a description of what is covered in the two statistics textbooks paired with this book plus a list of relevant further material that you can find on the book's Web site.

Methodological and statistical concepts covered in this chapter

5 percent significance level
Failing to reject the null hypothesis
Findings that are not statistically significant
Inferential statistics
One-tailed test of significance
Probability
Rejecting the null hypothesis
Reliability
Significance level (alpha level)
Statistical significance
Statistically significant findings
Two-tailed test of significance
Type I error
Type II error

We also discussed how to go about choosing a statistical test.

Diagnostic questions – check your understanding

1 What are *inferential statistics*?
2 What is a *statistically significant* difference?
3 What does it mean to say that the *5 percent significance level* was used in an analysis?
4 What is a *type I* error?
5 What is a *type II* error?

You can find the answers at the rear of the book (p. 276).

Statistics textbooks

∑ The books paired with *Designing and Reporting Experiments* have the following coverage:

Learning to use statistical tests in psychology – Greene and D'Oliveira
Greene and D'Oliveira have helpful coverage of how to choose a statistical test, which is a core theme of their book, with a section on this in each chapter. There is a handy decision tree inside the rear cover of the book and Chapter 3 describes how to use this. You will also find coverage of probability and statistical significance, including coverage of significance levels (pp. 16–17) and of two-tailed and one-tailed tests (p. 19). They discuss the assumptions of non-parametric tests in Chapter 5 and of parametric tests in Chapter 13 and compare them in Chapter 8.

SPSS survival manual – Pallant
Chapter 10 of Pallant gives excellent step-by-step coverage of how to go about choosing a statistical test, with a handy summary table at the end. She also repeats the key relevant points from this when covering which statistical technique to choose when comparing groups in the opening section to Part Five (pp. 201–5), where she also discusses type I and type II errors.

The *Designing and Reporting Experiments* Web site

In Section A of the book's Web site at **http://mcgraw-hill.co.uk/openup/harris/** you will find expanded coverage of Section 11.4 of this book on how to choose a statistical test.

12 Statistics: effect size and power

When we test for statistical significance, therefore, we end up either rejecting the null hypothesis or not rejecting it. Following this, we critically examine our experiment, looking for any flaws that might be rival explanations for our findings (see Chapter 5). The ultimate goal of this procedure is to be able to determine whether or not the IV has a causal effect on the DV. This is the traditional procedure by which psychological experimentation has worked.

However, some people argue that we should be asking a subtly different question of the data. That is, instead of asking whether there is or is not a causal effect of the IV on the DV, why not ask *how big* an effect there is? Does the evidence suggest that the IV affects the DV not at all, a small amount, a medium amount, or a large amount? For example, in our driving and music experiment, it would be useful to know that listening to music makes a big difference to driving performance rather than a small one.

The difference between the questions is subtle, but the implications are quite profound. For example, some of those who promote asking the "how big" question also advocate abandoning significance testing altogether. Some of the reasons for this may become clearer as you work through this chapter. For now, however, I want you to focus on two issues that arise from the "how big" question that are important to think about even when you are testing for statistical significance: the *effect size* of your IVs and the *power* of your experiments.

12.1 Effect size

There are statistics available to help us to answer the "how big" question. These statistics use the data to form estimates of the difference that the IV makes to scores on the DV. They are called statistics of **effect size**. For many years psychologists tended to ignore the issue of effect size and concentrated on whether or not an effect or a relationship was statistically significant. One psychologist in particular – Jacob Cohen – has been especially influential in making psychologists think more about the size of the effects that we report.

Put loosely, effect size is the magnitude of the difference between the conditions in an experiment. (With correlations the issue concerns the strength of the relationship between the variables.) In a well-designed experiment effect size is the difference that the IV makes to the scores on the DV. So, for example, in a well-designed, two-condition experiment of the sort discussed in Chapter 10, it is the difference between the scores when not exposed to the IV (the control condition) and when exposed to the IV (the experimental condition). Thus, everything else being equal, the bigger the impact that the IV has, the bigger will be the difference between the scores and the bigger will be the effect size.

Effect size can be measured in standard deviations. One of the advantages of this is that we can then compare effect sizes across different experiments. In order to help us to describe and compare effect sizes, Cohen (1988) suggested some guidelines for determining whether an effect is large, medium or small. He suggested that we consider an effect of at least 0.8 standard deviations to be a large effect size and one of around 0.5 standard deviations to be a medium effect size. An effect of 0.2 standard deviations, he suggested we regard as a small effect size.

You will learn more about effect size in your statistics course. The important thing to note for now, however, is that Cohen argued that in psychology effect sizes are typically small or medium, rather than large. This has important implications for the way that we should design our experiments and other studies. For, if many of the effects that we are interested in are likely to be small or at best medium, then we need to make sure that our studies have the *power* to detect them.

Summary of Section 12.1

1 Instead of asking whether there is or is not a causal effect of the IV on the DV, some authors suggest we should ask how big an effect there is.

2 There are statistics available to help us to answer the "how big" question. These are called statistics of effect size.

3 In an experiment, effect size is the relative magnitude of the difference between the conditions. It is often expressed in standard deviations.

4 Cohen has argued that in psychology effect sizes are typically small or medium, rather than large.

5 If the effects that we are interested in are likely to be small or medium, then we need to make sure that our studies have the power to detect small or medium effects.

12.2 Power

Let's imagine for a moment that we have had the privilege of seeing some of the tablets of stone on which are written the Laws of the Psychological Universe. These state that our IV – say, listening to music – has a particular size of effect – say medium – on driving performance. In this context, **power** is the capacity of our study to detect this. That is, power is the probability of *not* making a type II error, of *not* failing to find a difference that is actually out there in the psychological universe. The power of a statistical test is its capacity to *correctly* reject the null hypothesis.

Experiments vary in power. The more powerful the experiment is, the smaller are the effects it can detect. In this way, experiments to psychologists are like telescopes to astronomers. Very powerful telescopes can see even dim and distant stars. Likewise, very powerful experiments can detect even tiny effects of the IV on the DV. Low-power experiments are like low-power telescopes. Telescopes lacking in power may be unable to detect even fairly close and bright stars. Similarly, experiments lacking in power may be unable to detect even quite strong and important effects of the IV on the DV.

Power is measured on a scale from zero to one. With a power of zero, there is no chance of detecting an effect of the IV on the DV,

even if one exists. With a power of .25, there is a 25 percent chance of detecting a particular effect of the IV on the DV. This is not very likely and thus not very powerful. With a power of .80, there is an 80 percent chance of detecting such an effect of the IV on the DV and we are beginning to be in business. We need our designs to have at least this level of power if they are going to be worth running. So, how do we go about assessing and improving the power of our experiments?

12.2.1 Estimating power

Fortunately, we can estimate the power of our studies – including our experiments – before we run them. This enables us to check whether they will be powerful enough to detect the effects we are interested in. If they look like they will not be powerful enough we can then take steps to increase their power.

Power is influenced by a number of factors:

- the effect size;
- the significance level;
- whether the test is one- or two-tailed;
- the type of statistical test;
- the number of participants;
- whether the IV uses related or unrelated measures.

Perhaps you can see that the only one of these factors that we do not know for sure *before* running the experiment is the *effect size*. If we can make a reasonable guess about the likely effect size in our experiment, therefore, we can estimate the power of our experiment before running it. If necessary, we can then think about ways of increasing power.

There are two basic ways of estimating the likely effect size before running the experiment. The best way is to base the calculations on other experiments that have investigated the effects of this particular IV. This enables us to estimate the *likely* effect size of our IV. If such research is not available, however, we can still estimate how powerful our experiment would be if the IV had a given effect size. That is, we can assume that the effect size will be small or medium or large and calculate how likely our experiment would be to detect the effect under each of these circumstances. Again, we can then think about ways of increasing power if we find that the current design will be

unlikely to detect a small or medium effect. On the other hand, if we find that it has the power to detect a small effect, we know that it will also be able to detect a medium or a large effect.

These calculations, not surprisingly, are called **power calculations**. Typically, we use them to help us to estimate how many participants we will need for our experiments to have enough power (say, .80) to stand a reasonable chance of detecting the effect we are interested in. Alternatively, we can save some effort by looking up power in relevant *power tables* or by using one of the many Web sites that are available to calculate power for us. You may learn more about power and even about how to calculate it in your statistics course. If so, you will find that the results are salutary: it can be surprising to find out quite how many participants you will need to stand, say, an 80 percent chance of rejecting a false null hypothesis with an unrelated samples IV.

This leads us to the next question. How can we go about *increasing* the power of our experiments if we need to?

12.2.2 Increasing the power of our experiments

In theory, we could increase power in any of the following ways: by increasing the effect size, making the significance level less stringent, using a one-tailed test, employing a parametric test, increasing the number of participants, and using a related samples IV. In practice, however, some of these are more practical and desirable ways of increasing power than others.

Changing the effect size is, of course, beyond the powers of us mere mortals. As we saw in Chapter 11, making the significance level less extreme (e.g., using the 10 percent significance level instead of the 5 percent significance level) means that there is greater chance of rejecting the null hypothesis. However, it increases our chances of doing so when we should not (i.e., of making a type I error). We should therefore be cautious about altering this. Likewise, you should be cautious about using one-tailed tests as these limit the inferences that you can draw from your experiment (Section 11.5). You should use the most powerful statistical technique that your data allow (e.g., parametric rather than nonparametric tests where appropriate). The remaining routes of increasing the power of our experiments, therefore, are to run enough participants and to use related samples IVs whenever possible. Let us consider these.

One relatively straightforward way of increasing power is to run more participants, especially with unrelated samples IVs. You need as many as will enable you to detect an effect of the size that you suspect that the IV may have on the DV. For example, if listening to music has only a small effect on driving performance, you will need more participants to detect this small causal effect than if listening to music has a medium effect on driving performance. With too few participants you would fail to detect the effect and conclude (erroneously) that listening to music makes no difference to driving performance.

? SAQ 41 What type of error would you have made?

So, with unrelated samples in particular, increasing the number of participants can be an effective way of increasing the power of your experiment to the level you need. The larger the sample size, the greater the power (other things being equal). We will discuss this further in Section 13.1.1.

Using related samples increases power by reducing the background variation against which we have to assess the effect of the IV on the DV (Section 10.4). This is a very efficient way of increasing power. It is a bit like using a radio telescope that has a filter that cuts out a lot of the background "noise" – the irrelevant signals that can mask the signals from the less powerful stars. Filter this out and you can "see" the signals from these weaker stars more clearly. This is one of the reasons why I suggested earlier that you should try to use related samples IVs whenever possible. However, as you know, it is not always possible to use related samples for an IV (Section 10.4).

Summary of Section 12.2

1 Power, in this context, is the capacity of an experiment to detect a particular effect of the IV on the DV. It is thus the capacity to *correctly* reject the null hypothesis.

2 Experiments vary in power. The more powerful the experiment is, the smaller are the effects it can detect.

3 Experiments that lack power may be unable to detect even strong and important effects of the IV on the DV.

4 Power is measured on a scale from zero to one. The closer to one, the greater the power of the design and the better its chance of correctly rejecting the null hypothesis.

5 We can estimate the power of our studies before we run them. If we need to, we can then take steps to adjust their power. Doing this in advance helps us to avoid wasting time by, for example, running an experiment that has too little power to detect the effects we are interested in.

6 The principal ways in which we increase power are to run enough participants and to use related samples IVs whenever possible.

12.3 Effect size and power: reporting and interpreting findings

Let's recap. The more powerful the experiment, the smaller the effect sizes it can detect. If the design lacks power, however, we may not be able to detect even quite large effects. Therefore, our experimental designs can be insufficiently powerful to detect effects of our IV on our DV, even if such effects exist and even if the effect size is quite large. This is just like using a weak telescope to look for stars: if the telescope is weak, we will not be able to see even quite bright stars.

Of course, we would not conclude that the only stars that exist are those we can we see through a weak telescope! Likewise, it is important to make sure that our experiments are powerful enough to detect the effects we are searching for. Using our data to estimate effect size *as well as* testing for statistical significance assists us with this problem. Doing so helps us to assess whether the lack of statistical significance is a result of the IV having little or no effect on the DV (the question the experiment was designed to answer) or because the experiment lacked sufficient power.

As you can see, these issues have important implications for the studies we run and the ways in which we report and interpret our findings. In terms of design, it obviously helps to consider power from the outset, to think carefully about the size of the effect we are expecting and to adjust the design accordingly. However, what about the implications of effect size and power for reporting and interpreting our findings? The rest of this chapter has advice on this. (If you already know how to calculate power or effect size statistics, you can go straight to Section 12.3.2.)

12.3.1 Reporting for those who do not know how to calculate power or effect size statistics

Courses vary greatly in when they teach you about these issues. Some of you will be introduced to them from the very start of your careers as students of psychology. Others may find that you are taught about power and effect size only quite late on in your studies. Some of you will be taught how to calculate effect size statistics and even power. Others of you will be taught about these things conceptually but not taught the calculations. Nevertheless, most of you should have at least some awareness of these issues by the time you come to undertake project work (e.g., in the final year of a UK degree course). If you have not been taught how to calculate power or to obtain effect size statistics by that stage, you should at least be aware of how these issues affect the inferences that you can draw from and about statistically significant and statistically nonsignificant findings.

Here is some advice about what you can do in your reports if you have been taught about these things conceptually but not taught how to calculate effect size statistics or power.

In the RESULTS you can do the following, as I advised in Chapter 4:

1 Whenever possible, report the exact probabilities associated with the obtained values of your statistic, whether these are statistically significant or not.
2 Where appropriate, report standard deviations as well as means in tables of descriptive statistics.
3 If you know how, report appropriate confidence intervals such as the 95 percent confidence intervals around the mean, or for a difference between two means. You can do this regardless of whether or not the effects are statistically significant.

Reporting relevant confidence intervals is increasingly being encouraged in psychological research. There are a variety of reasons for this. Among them is the insight that such confidence intervals offer about the power of your experiment. To put it simply, the more powerful the study (other things being equal), the narrower will be the confidence interval. This will help you to interpret results in which the null hypothesis has not been rejected. With a low-power study, the confidence interval will be wide and will encompass everything from zero effects to quite large effects of the IV. With a high-power study,

the same confidence interval will be narrow and will encompass zero to very small effects only.

Those of you using and reporting the correlation coefficient Pearson's *r* might also consider reporting and commenting on an easily computed statistic of effect size. If you *square* the value of *r* (i.e., calculate the statistic called r^2), this will tell you how much of the variance is shared between the two variables in the correlation. This is a measure of effect size. It is useful to do this, as it will help you to avoid giving precedence to the statistical significance of the coefficient rather than to the size of the relationship. For, with enough participants, even correlations as small as $r = .2$ or even $r = .1$, can become statistically significant at the 5 percent significance level. Yet these are very small relationships. An $r = .2$, for instance, gives $r^2 = .04$, which indicates that only 4 percent of the variance is shared. So, think about calculating r^2 when you use Pearson's *r*.

4

One thing to avoid at all costs – especially in the DISCUSSION – is to confuse statistical significance with the size of the effect. If you have been able to follow the drift of the discussion in this chapter, you will perhaps see how the problem arises. If the experiment or study is powerful enough, *any* difference between your conditions will be statistically significant, no matter how small it is numerically and trivial it is psychologically. (Many of you realize this intuitively when you write in your reports that "our findings would have been significant if we had had more participants". It is precisely because this is *inevitably* true that – unless you defend it sensibly – this remark tends to evoke the scorn and wrath of your marker.) A very powerful experiment may well tell you that an effect is statistically significant even if the IV has only a relatively trivial effect on the DV. Thinking about effect size *in addition* to statistical significance helps you to avoid this problem. Doing so helps you to assess whether the effect is large enough to be potentially interesting psychologically. Remember that with a very powerful experiment it is possible for an effect to be highly significant *statistically* but for the effect size to be so small that the effect of the IV on the DV is trivial, both numerically and psychologically.

Don't panic if you don't yet know how to produce or calculate confidence intervals or statistics of effect size! Even without them you can still make some efforts to address effect size in your report. *Look at the sample size*. If the experiment has been well designed in other respects, the sample size is small and yet the effect is statistically significant (especially on an unrelated samples IV), then this suggests

that the IV has a pretty large effect size. However, if the sample is large, then the results are ambiguous, for even trivial effects may be statistically significant with large samples. If the experiment has been well designed in other respects, the sample size is small and the effect is not statistically significant (especially on an unrelated samples IV), then the results are ambiguous, for even large effects may not be statistically significant with small samples. However, if the sample is large, then the failure to reject the null hypothesis suggests the absence of an effect of much size or psychological importance. (Of course, this depends on just how large the sample size is.)

It is very easy to confuse statistical significance and effect size, and people often do. So, do not worry if this discussion doesn't fall into place at once. These issues require thought and effort before they become clear. You may need to reread this chapter and also your statistics textbook several times before the penny even begins to drop.

12.3.2 Reporting for those who have been taught how to calculate power or effect size statistics

Obviously, if you are taught how to do even rudimentary power calculations or to use power tables or Web sites, then you should include these in the METHOD of your report. Advice on how to do this is in Section 3.6.3.

If you know how to calculate confidence intervals or how to get a statistical software package to produce these, then include them in the RESULTS (see Section 4.6.10). Likewise, once you know how to calculate or produce effect size statistics and understand what they mean, you will be able to report relevant statistics of effect size *as well as* the obtained value of your statistic and its associated probability. To give you an idea of how you might do this, here is the paragraph describing the principal findings from the mnemonic experiment, this time with an effect size statistic added. This statistic is known as *partial eta squared* (η^2) and is one of the effect size statistics available when using ANOVA.

5

> The data in Table 1 were analysed using 2 × 2 ANOVA for mixed designs, with imageability (easily imaged or hard to image) as the related samples variable and instruction (mnemonic or no mnemonic) as the unrelated samples variable. There was a statistically significant main effect of instruction, $F(1, 38) = 7.20$, $p = .01$, with those in the

mnemonic group recalling more items overall than did those in the no mnemonic group (M = 15.65, SD = 3.97; M = 12.40, SD = 3.74, respectively). Partial η^2 = .16, indicating that 16% of the overall variance was attributable to this manipulation. There was also a statistically significant main effect of imageability, $F(1, 38) = 145.22$, $p < .001$, with more items from the easily imaged list being recalled than from the hard-to-image list (M = 15.98, SD = 4.12; M = 12.08, SD = 4.48, respectively). Partial η^2 = .79, indicating that 79% of the over-all variance was attributable to this manipulation. However, these main effects were qualified by the significant Instruction X Imageability interaction, $F(1, 38) = 11.55$, $p = .002$. Partial η^2 = .23, indicating that 23% of the overall variance was attributable to the interaction between the variables. Figure 1 displays this interaction.

You should use these statistics in the DISCUSSION to help you to answer the following questions:

1 If an effect is statistically significant, what do the effect size statistics tell us about the likely size of the effect? Is the effect big enough to be of psychological interest and importance?
2 If an effect is not statistically significant, what do the confidence intervals tell us about power? Are the 95 percent confidence intervals broad (suggesting that the analysis lacked power) or narrow (suggesting that it did not)? What do the effect size statistics indicate about the likely size of the effect? If these indicate that the effect is not small, this suggests that the analysis lacked power. If you conclude that the experiment lacked the power to detect the effects, suggest what steps might be taken in future to run studies with sufficient power.

Perhaps now you can see why some researchers have called for the abandonment of significance testing altogether. If we report statistics of effect size together with the exact probability associated with the obtained value of our statistic, so the argument runs, then at best statistical significance is redundant and at worst it is downright misleading. Pursuing this debate is beyond the scope of this book. However, you may well discuss these issues further in your statistics course.

Summary of Section 12.3

1 There are some things that you can do about these issues in your reports even if you have not been taught how to calculate power or to obtain effect size statistics.

2 These include reporting the exact probability associated with your obtained statistic and, once you know how, reporting appropriate confidence intervals.

3 Whatever you do, make sure that you do not confuse statistical significance with effect size. A very powerful experiment may well tell you that an effect is statistically significant even if the IV has only a relatively trivial effect on the DV. On the other hand, a large effect may not be found to be statistically significant if the design lacks power.

4 Even without statistics of effect size, looking at the sample size can help you to make some inferences about effect size.

5 If you do any power calculations, report this in the METHOD. You may also be taught to generate other statistics indicative of effect size and power. If so, report these statistics in the RESULTS and use them to help you to interpret your findings in the DISCUSSION.

Consolidating your learning

This closing section is designed to help you check your understanding and determine what you might need to study further. You can use the list of methodological and statistical concepts to check which ones you are happy with and which you may need to double-check. You can use this list to find the relevant terms in the chapter and also in other textbooks of design or statistics. The diagnostic questions provide a further test of your understanding. Finally, there is a description of what is covered in the two statistics textbooks paired with this book plus a list of relevant further material that you can find on the book's Web site.

Methodological and statistical concepts covered in this chapter

Effect size
Power calculations
Statistical power

Diagnostic questions – check your understanding

1 What do we mean by *effect size*?
2 What is the *power* of an experiment?
3 What does it mean to say that an experiment has a power of .80?
4 What are *power calculations*?
5 How practically might we *increase* the power of an experiment if we did not want to change the significance level and could not use a one-tailed test of significance?
6 What percentage of the variance is shared by two variables where $r = .50$?

You can find the answers at the rear of the book (p. 277).

Statistics textbooks

∑ The books paired with *Designing and Reporting Experiments* have the following coverage:

SPSS survival manual – Pallant
Pallant covers power and effect size in the introduction to Part 5. Helpfully, she also describes how to obtain effect size statistics from SPSS for most of the tests included in the book. For example, partial eta squared (the effect size statistic referred to in this chapter) is described in the introduction to Part 5 and in Chapters 18 and 19.

The *Designing and Reporting Experiments* Web site

In Section B of the book's Web site at **http://mcgraw-hill.co.uk/openup/harris/** you will find coverage of some issues to consider when reporting statistics of effect size.

13 More advanced experimental design

Once you have mastered the basics, you can start to think about designing more complex experiments. This will enable you to ask more subtle questions of the psychological universe, such as whether variables *interact* with each other. There are two basic ways of extending experiments:

1 You can use an IV with more than two levels.
2 You can manipulate more than one IV simultaneously.

You can do either of these things or both at the same time.

13.1 Extending the number of levels on the independent variable

As you know, we need not restrict ourselves simply to manipulating the *presence* versus the *absence* of the suspected causal variable. Experiments using IVs with three or more levels are very common in psychology and are not hard to design. As with any IV, these can employ either unrelated or related samples. The principal design issues that arise have to do with using enough participants with unrelated samples to ensure sufficient power or being able to adequately control for order effects with related samples.

13.1.1 Unrelated samples IVs

Having adequate numbers of participants to ensure sufficient power is an issue for *all* experimental designs, but it is an especially pressing issue with unrelated samples because it is so easy to design a study that lacks power. As discussed in Section 12.2, power in this context refers to the capacity of your experiment to detect an effect of the IV on the DV where one exists. The *more* participants you use, all things being equal, the *more* likely you will be to detect any impact of the IV on the DV. Indeed, unless the IV has a pretty dramatic impact on the DV (e.g., listening to music *massively* impairs or improves driving performance), designs using unrelated samples with small numbers of participants will be useless.

How many participants do you need? The number of participants you need depends on a number of factors, including how strong you estimate the effect of the IV on the DV will be (Section 12.2.2). Sadly, therefore, there is no simple answer to this question. (It really is just like asking how long is a piece of string.) Obviously, the ideal way of estimating this number is through *power calculations* (Section 12.2.1). However, this is of little help to those of you who have no idea how to obtain such calculations. For you, the only sensible answer I can offer is that this number is almost certainly going to be much larger than the one you first thought of! Thinking in terms of 10–15 participants *per condition* is unlikely to give you much chance of rejecting a false null hypothesis. Even double these numbers may be insufficient. As an incredibly rough rule of thumb, with a number plucked out of the sky, look to use no fewer than 20 per condition and try to use substantially more than this if you can. You therefore need to think very carefully when considering extending the number of levels on an IV with unrelated samples. Make sure that you have the time and the resources to get reasonable numbers of participants in each condition. When in doubt, keep it simple!

Particularly early on in your career, you may *have* to use fewer participants than this (e.g., for reasons of time or problems with access to people who can take part in the experiment). Under these circumstances, remember two things. First, if you are playing a role in the design of the experiment (as opposed to having it designed for you by a tutor) then make sure that the IV *has* to use unrelated samples. Moreover, think carefully *before* extending the levels on this unrelated samples IV beyond two or three. Second, if your findings are

not statistically significant, bear in mind that a potential explanation for this is that the design was not powerful enough to detect the effect, rather than that there was no effect to detect.

Later on (for example, in final-year projects in the UK), this is likely to be one of the issues on which you are assessed. When using IVs with unrelated samples you need to have sufficient numbers of participants in each of the conditions, so bear in mind when designing the experiment the number of participants likely to be available. However, do not overreact to this problem by using *too many* participants. The number of participants you run should be optimum, as determined by power calculations or, failing that, reasonable guesswork.

Summary of Section 13.1.1

1 With unrelated samples the principal design issue arising from IVs with more than two levels is to use enough participants in each condition to ensure sufficient power.

2 This means that you will have to think carefully when considering extending the number of levels on an unrelated samples IV. Will you have the time and resources to get enough participants in each condition?

3 Do not overreact to this issue by using too many participants. The number of participants should be optimum – either as determined by power calculations (preferably) or by reasonable guesswork if you cannot do the calculations.

4 With findings that are not statistically significant, examine whether this is because the experiment lacked sufficient power to detect an effect.

13.1.2 Related samples IVs

As we discussed in Chapter 10, numbers of participants is much less of a problem when using *related* samples. This is because of the greater power of related samples (Section 12.2.2). Here, instead, the principal problem is usually *order effects*. Exactly the same issues arise as we considered in Section 10.4.3. You cannot eliminate order effects; you therefore need to ensure that the effects due to order contribute approximately equally to each of the experimental conditions. As

Table 13.1
All Possible Combinations of Four Conditions (A, B, C and D)

ABCD	BACD	CBAD	DBCA	ACDB	BCDA
CADB	DCAB	ACBD	BCAD	CABD	ABDC
DCBA	BADC	CBDA	DBAC	ADBC	CBDA
BDAC	DABC	ADCB	BDCA	CDAB	DACB

Note. Use this table to help you to counterbalance any experiment in which you have a four-level, related samples IV.

before, the ideal way of achieving this is to *counterbalance* the design. With three conditions, there are six different combinations (given by three factorial, $3! = 3 \times 2 \times 1 = 6$). With four conditions, there are 24 different combinations ($4! = 4 \times 3 \times 2 \times 1 = 24$), as you can see in Table 13.1. You can use this table to help you to counterbalance any experiment in which you have a four-level, related samples IV – it lists for you *all* of the possible combinations.

With five conditions there are actually 120 different combinations. This means, in effect, that counterbalancing is possible for most of us only up to IVs with *four* levels. After that we need too many participants and will have to use alternative controls for order.

? SAQ 42 Suppose that you wanted to run an experiment with three conditions using related samples.

(a) Which, if any, of the following numbers of participants would you need to run to ensure the design was counterbalanced?

3 5 6 10 12 15 18 20 22 25

(b) Which, if any, of the numbers you have chosen would it be better to use?
(c) What is the next highest number you could use to ensure that the design was counterbalanced?

? SAQ 43 Suppose that you wanted to run an experiment with *four* conditions using related samples.

(a) What would be the minimum number of participants you could run to ensure that this experiment was counterbalanced and why?

(b) If you considered that this minimum number was unlikely to be sufficient to detect the impact of the IV on the DV, what would be the next highest number of participants that you could run to ensure that it was counterbalanced?

If for some reason you cannot counterbalance to control for the effects of order, then you might try partially controlling for these effects instead. For example, the first column in Table 13.1 contains four orders in which every condition appears in each ordinal position (first, second, third, fourth) and is followed by and preceded by every other condition once only. You could use this to partially control for order in an experiment involving fewer than 24 participants. (For example, if you had 12 participants you could randomly assign each of the 4 orders to groups of 3 participants using randomization without replacement – see Appendix 2.)

When even this solution is not possible, then turn to that good old standby, *randomization*. Just as before, however, remember that randomization does *not* guarantee that the order effects will be spread around equally, and also that it is likely to be more successful in achieving this (though again not inevitably so) the more participants you have.

Finally, remember that controls for order are *only* useful where there are no *carry-over* effects. Just as before, if you suspect that doing one of the conditions before another will have a permanent or disproportionate effect on performance, then do not use related samples for that IV.

? SAQ 44 Those of you who relish a challenge might now like to go back to SAQ 24 and work out how you could control for order effects in the two designs using related samples, (a) and (e). State what the minimum number of participants would be for these experiments, given the controls you have chosen to employ.

Summary of Section 13.1.2

1 With related samples IVs with more than two levels the principal design issue is controlling for order effects.

2 With IVs with three or four levels it may be possible to control for order by counterbalancing. (Table 13.1 lists all the possible

combinations of a four-level IV to help you counterbalance such a design.)

3 With more than four levels you are more likely to control for order partially or by randomizing.

4 Controls for order are only useful when there are no carry-over effects. If doing one of the conditions before another will have a lingering, permanent or disproportionate effect on participants' responses, then related samples should not be used for that IV.

13.2 Experimental designs with two or more independent variables

Experiments involving more than one IV are very common in psychology. Such experiments enable us to test more sophisticated ideas about the variables in the psychological universe, such as whether the variables *interact* with each other.

For example, rather than simply concerning ourselves with the effects of having the music on or off on driving performance, we might want to find out whether this affects those who have been drinking alcohol more than those who have not been drinking alcohol. In order to do this we could design an experiment in which we simultaneously manipulated *both* the IV *music* and the IV *alcohol*. How would we do this? Well, we could turn to our pool of participants and assign half of them randomly to a music on condition, and half to a music off condition. Of those in the music on condition, we could assign half randomly to drink a standard amount of alcohol, with the remainder remaining sober. Likewise with the music off condition, we could ask half of our participants to drink the standard amount of alcohol, with the other half remaining sober (Table 13.2). We would thus have manipulated *two* independent variables, (music and alcohol) with two levels on each (music on or music off, and alcohol or no alcohol). This is quite a common design in psychology. It is sometimes referred to as a **two by two** or **2 × 2** design.

Of course, we need not restrict ourselves to manipulating the *presence* versus the *absence* of the suspected causal variables. We could, for instance, vary the *amount* of alcohol that our participants consumed, so that one group drank no alcohol, another drank the equivalent of one glass of wine, another the equivalent of two glasses of wine, and another the equivalent of three glasses of wine. At the

Table 13.2
An Experimental Design with Two Independent Variables, Two Levels on Each

	IV: music	
IV: alcohol	Level 1: off	Level 2: on
Level 1: no	Condition 1	Condition 2
Level 2: yes	Condition 3	Condition 4

same time we would have half of the participants driving with the music on, the other half with the music off. Such a design appears in Table 13.3. On the other hand, we could simultaneously vary the *volume* at which the music was played – with one group listening to the music at low volume, another at medium volume, and a third at high volume. Such a design appears in Table 13.4.

In principle, there is no limit to the number of independent variables you can manipulate simultaneously in an experiment. Neither is there any restriction over whether these variables employ related or unrelated samples. For example, in the first version of the above experiment the variables could *both* be unrelated (with different participants in conditions 1–4) or *both* related (with the same participants in conditions 1–4, albeit with a suitable time lag between the alcohol and the no alcohol conditions). Or we could have a *combination* of related and unrelated samples (for instance, the music variable could employ *related* samples, the alcohol variable, *unrelated* samples).

Table 13.3
An Experimental Design with Two Independent Variables, One with Two Levels, the Other with Four Levels

	IV: music	
IV: alcohol[a]	Level 1: off	Level 2: on
Level 1: 0 glass	Condition 1	Condition 2
Level 2: 1 glass	Condition 3	Condition 4
Level 3: 2 glasses	Condition 5	Condition 6
Level 4: 3 glasses	Condition 7	Condition 8

Note. [a] Glasses of wine, where each glass contains 1 standard unit of alcohol.

Table 13.4
An Experimental Design with Two Independent Variables, Four Levels on Each

	IV: music			
IV: alcohol[a]	Level 1: off	Level 2: low volume	Level 3: medium volume	Level 4: high volume
Level 1: 0 glass	Cond 1	Cond 2	Cond 3	Cond 4
Level 2: 1 glass	Cond 5	Cond 6	Cond 7	Cond 8
Level 3: 2 glasses	Cond 9	Cond 10	Cond 11	Cond 12
Level 4: 3 glasses	Cond 13	Cond 14	Cond 15	Cond 16

Note. Cond = Condition.
[a] Glasses of wine, where each glass contains 1 standard unit of alcohol.

Given this flexibility, it is important to know how to design and run studies involving more than one IV. It is also important to know how to refer to such designs in the DESIGN and RESULTS. The next section describes two ways of labelling such designs.

Summary of Section 13.2

1 It is perfectly possible to design experiments in which we manipulate more than one IV at the same time.

2 In experiments involving more than one IV, each can have two or more levels. The IVs can be exclusively unrelated samples or exclusively related samples IVs or any combination of unrelated and related samples IVs.

13.3 Labelling designs that have two or more independent variables

One way of labelling these designs is to describe them in terms of a combination of the *number* and *nature* of their IVs. So, for instance, the experiment in Table 13.2 would be called a "two-way" experiment, because it has two IVs. If both IVs used unrelated samples, the experiment would have a "two-way, unrelated samples" design. If

both IVs used related samples, it would have a "two-way, related samples" design. If one of the IVs used unrelated samples and one used related samples, the experiment would have a "two-way, mixed" design, for designs involving both types of IV are described as having **mixed designs**. Using this convention, the designs that we talked about in Chapter 10 were *one-way* designs.

So, to label your design correctly using this method, first count how many IVs there are in your design. If there is one IV, it will be a *one-way* design. If there are two, it will be a *two-way* design. With three it will be a *three-way* design, and so on. Once you have determined this, then work out for each IV whether it uses related samples or unrelated samples. If all the IVs use related samples, then it is a *related samples* design. If all the IVs use unrelated samples, then it is an *unrelated samples* design. If there is at least one of each type of IV, then it is a *mixed* design. The logic extends to "five-way" experiments, "six-way" experiments, and so on. However, my advice is that you restrict yourself to experiments with a maximum of three IVs (Section 13.6).

If you have been taught to use different labels for your IVs than the ones that I am using in this book, then use these to label the design. For example, you might call your design a "two-way, within-participants" or a "three-way, between-participants", and so on. If so, just substitute the label that you have been taught to use for the ones I am using.

When writing your report you will need to specify the number of levels on each of the IVs and what these levels were. With mixed designs, you will need to make clear which of the IVs used related samples and which used unrelated samples. (See Chapter 3 for an example.)

Another way of labelling these designs is to refer to them by describing the number of levels on each of the IVs. For example, the design in Table 13.2 could be described as a "2 × 2 related samples" design if you had used only related samples IVs. If you had used only unrelated samples IVs, it could be described as a "2 × 2 unrelated samples" design. If you had used one of each type of IV it could be described as a "2 × 2 mixed" design. Likewise, a three-way design involving one unrelated samples IV and two related samples IVs with three, three, and four levels respectively could be referred to as a "3 × 3 × 4 mixed" design, and so on.

Again, when writing your report you will need to make clear which IVs had which number of levels and what these levels were. With

mixed designs, you will need to specify which of the IVs used related samples and which used unrelated samples.

? SAQ 45 Below you will find a list of studies that have more than one IV. Go through each in turn and attempt to name these designs using *both* of the conventions outlined above.

(a) An accident researcher is interested in the effects of different levels of alcohol (0, 2, 4, or 8 standard unit measures of wine) on driving performance. Over the course of several weeks, she varies the quantity of alcohol given to a group of participants on different occasions. She then examines whether the impact of the alcohol varies with the person's gender and the length of time since they passed their driving test (passed in the previous 6 months, in the previous 7–18 months, or more than 18 months previously).

(b) An occupational psychologist is interested in the effects of different types of physical stressor (in this case heat, noise, or light) and the nature of the task (demanding or undemanding) on blood pressure. He exposes different groups of participants to the different types of stress, and measures their performance on both types of task.

(c) A psychologist interested in personality wishes to examine the effects of gender and extraversion (extravert or introvert) on public speaking performance.

(d) Another accident researcher is interested in the effect of listening to in-car audio on driving performance. Using a driving simulator, she varies the driving conditions (easy, moderate, or difficult), the nature of the material listened to (talk or music), and the volume at which it is played (low, moderate, or loud), on the driving performance of a group of experienced drivers.

Note that sometimes you will find designs with more than one IV referred to as *factorial designs*, and also come across references to *repeated measures* or to *split-plot designs*. **Factorial designs** are ones that use all the possible combinations of the levels of the different IVs. The designs in Tables 13.2–13.4 are all factorial designs. **Repeated measures** is another name given to designs employing (usually only) IVs with related samples. **Split-plot** is an alternative term for a mixed design that stems from the origins of this research design in

agricultural research. Other names may also be used. However, you should never come across (or use) the term "one-way, mixed design".

? SAQ 46 Why should you never use the term "one-way, mixed design"?

Summary of Section 13.3

1 You must be able to label accurately designs involving more than one IV.

2 In your report you must make clear for each IV in your experiment whether it used a related or unrelated samples IV, how many levels it had and what these levels were.

3 You can do this succinctly with a basic design statement.

4 This section described two ways of labelling designs that you could use in your report. One involves specifying the number of IVs in the basic design statement (e.g., two-way, three-way, and so on). The other involves specifying the number of levels on each IV in the basic design statement (e.g., 2×2, $2 \times 3 \times 3$, and so on).

5 Designs involving a combination of unrelated and related samples IVs are called mixed designs.

6 Although these methods have been illustrated using the labels for IVs employed in this book (unrelated and related samples IVs) other terms that you may use for these IVs can easily be substituted for these labels in the basic design statement.

13.4 Main effects of independent variables

With designs involving two or more IVs we can test for *interactions* between the different IVs, as well as for *main effects* of each IV. Testing for interactions enables us to test more sophisticated ideas about the ways in which causal variables operate in The Psychological Universe. First, however, let us consider main effects.

The **main effect** of an IV is the effect of that IV on the DV, *ignoring* the other IVs in the experiment. For example, in the experiment in Table 13.2, a significant main effect of alcohol would indicate that

driving performance with alcohol differed from driving performance without alcohol. A significant main effect of music, on the other hand, would indicate that driving performance when listening to music differed from driving performance when not listening to music.

One way of thinking about experiments employing more than one IV is to think of them as if they were several experiments being run simultaneously. The number of experiments being run corresponds to the number of separate IVs. There is potentially a significant main effect for every IV in the experiment, irrespective of whether it uses related or unrelated samples. For example, if we ran an experiment involving three IVs, this would be like running *three* experiments simultaneously. We could thus test for three main effects – one for each of the IVs in the design (see Table 13.5 for the main effects in experiments with two and three IVs).

? SAQ 47 How many significant *main* effects could we find in each of the experiments in SAQ 45?

You can find more about main effects in Section C of the Web site that accompanies this book at **http://mcgraw-hill.co.uk/openup/harris/**.

13.5 Statistical interactions

When we combine IVs in this way in an experiment, however, we get several bonuses. One of the most important of these is that we can test whether our IVs *interact* with each other. When IVs **interact** the effect of one IV is *different* at the levels of the other IV or IVs. For example, in Figure 13.1 you can see that drinking alcohol makes a bigger difference to performance when participants listen to music than it does when they do not listen to music. That is, the *difference* between driving performance with alcohol and with no alcohol is bigger at one level of the music IV (music) than at the other level of the music IV (no music). Therefore, the IVs interact.

Figure 13.2 contains another example of an interaction. Here you can see that the IVs interact to such an extent that the effects of alcohol are actually *opposite* in direction at the levels of the music IV. Here driving is worse among those who have drunk alcohol when they do not listen to music but better among those who have drunk alcohol when they *do* listen to music. (Obviously, I have made these

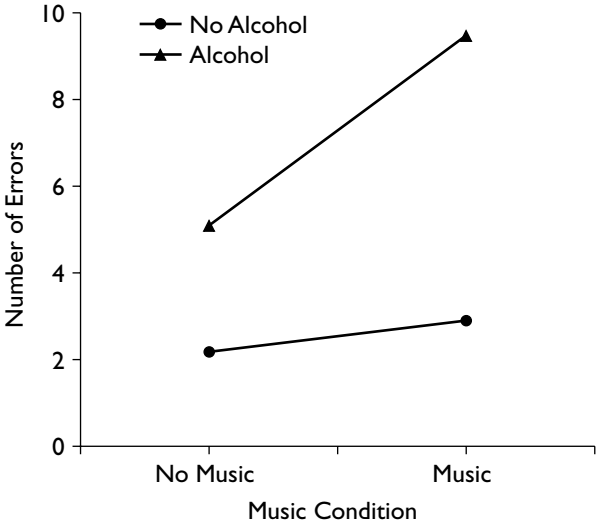

Figure 13.1. Line graph displaying one type of interaction between two IVs.

effects up!) Again, the IVs interact. That is, the effect of one IV is *different* at the levels of the other.

Both of these are examples of interactions. Had the IVs *not* interacted then, when graphed, the data would have looked more like Figure 13.3. Here the difference between driving performance with alcohol and with no alcohol is more or less the same when participants listen to music as it is when they do not listen to music. Therefore, the IVs do *not* interact.

As a student you can sometimes struggle to understand statistical interactions. Yet, when you first start designing your own studies in psychology, you are in fact naturally very aware of the possibility that variables can interact. Remember those first studies in which you wanted to include all the variables that you could think of? From ethnic background, social class, time of day, gender of participant, gender of experimenter, temperature in the room and what the participants had for breakfast, to how much coffee they had drunk? This reflects your natural, intuitive awareness that these variables may well make a difference to the effects. That is, you were concerned that the effect of the variables of interest may be changed – accentuated, reduced or even eliminated – in the presence of other variables. This is what an interaction tells us – that the effects of an IV are different – bigger, smaller or even eliminated – at the levels of another IV.

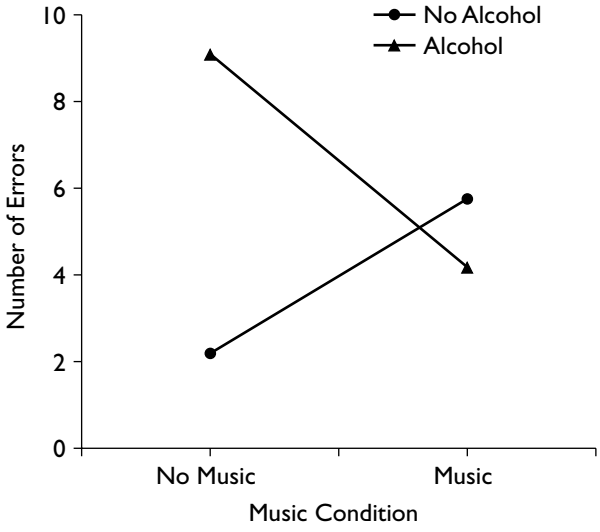

Figure 13.2. Line graph displaying an alternative type of interaction between two IVs.

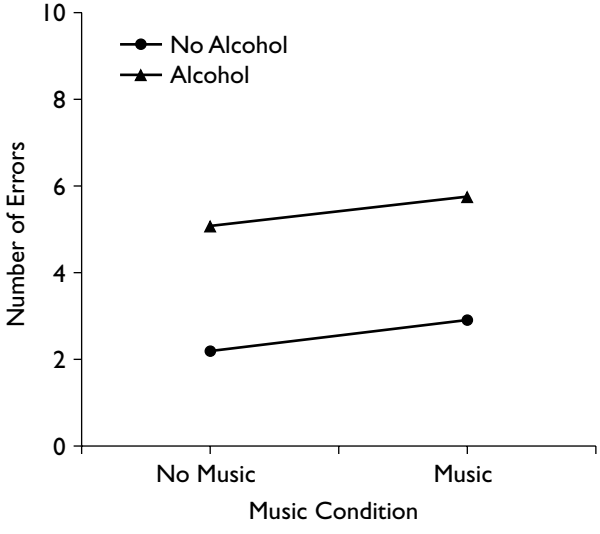

Figure 13.3. Line graph displaying no interaction between the two IVs.

A statistically significant interaction therefore tells us that the effects of one IV are *inconsistent* at the levels of other IVs. Although it might sound odd to be interested in inconsistencies in the effects of our IVs, we are often very interested in evidence that our IVs interact in this way. You can find out more about interactions in Section C of the book's Web site at **http://mcgraw-hill.co.uk/openup/harris/**. You will also hear more about these effects on your statistics course.

13.6 Analysing designs involving two or more IVs

Designs involving the simultaneous manipulation of two or more IVs are commonly analysed using the statistical test known as **analysis of variance** (ANOVA). You will certainly learn more about ANOVA on your statistics course. With ANOVA you can simultaneously test *all* of the main effects and interactions in your experiment for statistical significance. To help you with this, Table 13.5 lists all of the effects that you can find in designs using two and three IVs. Use this table to check against your statistical output. It will help you to keep track of the effects that you can expect to find in the analysis. Any or all of the effects listed there can be found to be statistically significant in a particular analysis.

Table 13.5
The Potential Main Effects and Interactions in Designs Employing Two and Three Independent Variables

Experiment with 2 IVs	Potential main effects:	IV1
		IV2
	Potential two-way interaction:	IV1 × IV2
Experiment with 3 IVs	Potential main effects:	IV1
		IV2
		IV3
	Potential two-way interactions:	IV1 × IV2
		IV1 × IV3
		IV2 × IV3
	Potential three-way interaction:	IV1 × IV2 × IV3

Note. Section C of the book's Web site contains an expanded version of this table showing the potential main effects and interactions in an experiment with four IVs.

Note that you do not need to know the number of levels or whether the IVs use unrelated or related samples in order to be able to use Table 13.5. You *do*, however, need to know these things in order to analyse the data appropriately. For this reason it is essential that you specify *which* particular version of ANOVA you used in your analysis. So make sure that you include an accurate and sufficiently detailed statement about which ANOVA you used in the RESULTS (see Section 4.6.10).

Finally, be careful about how many IVs you use in an experiment. My advice is to restrict yourself to three at the most, unless you have lots of experience or will receive very close supervision from a tutor. One reason for this is that in such designs controlling for order, matching materials, or obtaining enough participants is difficult. The main reason, however, is that interpreting three-way interactions is hard and interpreting interactions above and beyond three-way interactions (e.g., four-way interactions) is *so* hard that it can make your brain start to feel like it is slowly dissolving.

13.7 Graphing statistical interactions

Graphing significant interactions is a *very* useful aid to interpreting them. This was evident even for the 2×2 designs that I discussed in Section 13.5. It is even more evident when you have more levels on your IVs or more IVs involved in the interaction (or both).

I used **line graphs** in Figures 13.1–13.3 to illustrate the effects. The principal reason for this is that, because they represent inconsistencies in effects, interactions are revealed by deviations from the parallel. This is easy to see with a line graph. Technically, however, I should have used a *bar graph* for these examples, as the IV on the horizontal axis (music) is not quantitative. That is, the conditions *no music* and *music* are absolute, categorical states with no gradations between them. Yet there is a line on the horizontal axis for points between them and this is meaningless. Had I put alcohol on the horizontal axis this would have made sense, as there are sensible gradations between 0 units and 2 units of alcohol. Figure 13.4 is a *bar graph* showing the same interaction as in Figure 13.1. Take the advice of your tutor on which type of graph to use. Whichever you do use, get into the habit of graphing statistically significant interactions to help you to interpret them, and include figures displaying the key ones in your report

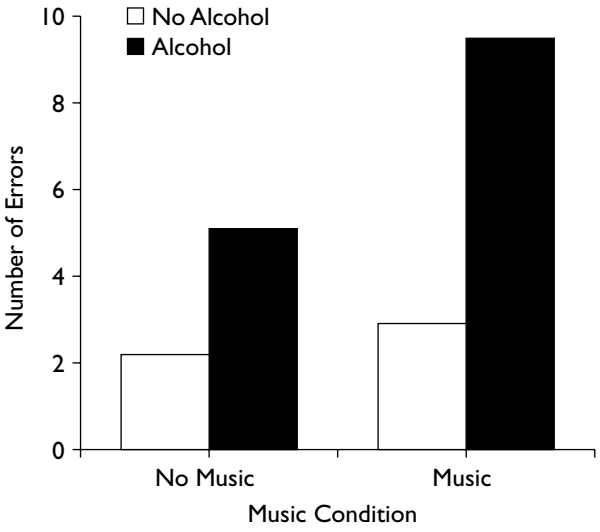

Figure 13.4. Bar graph displaying the same effects as in Figure 13.1.

(see Section 8.5). (You can find advice on how to graph three-way interactions in Section C of the book's Web site at **http://mcgraw-hill.co.uk/openup/harris/**)

Summary of Sections 13.4–13.7

1 In experiments involving more than one IV we can differentiate and test for the main effects of the different IVs and for interactions between the different IVs.

2 The main effect of an IV is the effect of that IV disregarding the other IVs in the experiment. There is potentially a significant main effect for each IV in the experiment.

3 An interaction occurs when the effects of one IV are different at the different levels of another IV.

4 It is difficult to interpret interactions between three or more IVs. So, as a general rule you would be well advised to limit the number of IVs that you manipulate in an experiment.

5 You should get into the habit of graphing statistically significant interactions and include graphs of the key ones in your report.

13.8 Watch out for "IVs" that are not true independent variables

Sometimes you will include in your experiments variables that look like IVs but are not true independent variables. These are variables like gender (male or female), personality variables (e.g., optimists or pessimists; people who are high, medium or low in trait anxiety), or differences in attitudes or behaviour (e.g., being for or against abortion; being a smoker or a non-smoker). Variables such as these are not *truly* independent variables because we have not been able to assign participants randomly to the different levels of the variable. That is, participants arrive at our study *already* assigned to the levels of these variables. For example, they come to us *already* as women, or against abortion, or smokers. Thus we do not determine their assignment to groups – it is predetermined.

Take the driving and music experiment. Imagine that instead of alcohol we are interested in gender differences in driving performance and the possibility that gender interacts with listening to music. We take our participants and allocate them randomly to the music on and the music off conditions. We take our participants and allocate them randomly to the women and men conditions. Er, except we cannot. Every time we reach for a human being who is neuter we find that they are already either male or female. Despite our best efforts, we are simply unable as experimenters to control who will be male and who will be female in our experiment. The allocation of participants to the levels of our variable *gender* is predetermined. Whenever we find that this is the case – whenever we are unable as experimenters to allocate our participants randomly to conditions – then we do not have a *truly* independent variable. (Do not confuse this with our control over *which* men and women participate in our experiment. The issue is our inability to control which sex any given participant gets given, not our ability to control who takes part in the experiment.)

This does not mean that we should avoid testing for gender differences or for differences in response between people classified in other ways that are predetermined. On the contrary, we often want to know a great deal about what other responses and variables are associated with being male or female, politically committed or apolitical, being of average or above average IQ, for or against genetic engineering, and so on. The important point is that, even though these variables look like IVs, we must remember when reporting the study and

interpreting the findings involving these variables that they are not IVs in the strict sense and that our conclusions must therefore be qualified and cautious.

Why qualified? Because, despite appearances, they are *not* experimental but correlational (see Section 9.2). This means that we cannot assume that they are the *causal* variables when we find changes on the DV associated with levels of this "IV".

Let's take an example. Suppose that you and I are interested in the impact of laughter on physical health. One approach to this would be to measure how much people laugh during a 5-minute interaction and relate this to a measure of physical health, say, how often they visit the doctor in the following 6 months. We could even form two levels on this "IV" by differentiating those who laughed *more* than average (high laughter) from those who laughed *less* than average (low laughter). (Let's do this by splitting them at the median for laughter, putting those above the median into the high laughter group and those below the median into the low laughter group.) Imagine that we find that there is indeed a difference, such that the above average laughter group pay fewer visits to the doctors. Would we be able to conclude that laughter *caused* this difference in visits? No, we would not. There are countless ways in which people who laugh a lot differ from those who laugh little. Take a minute or two to think of reasons why some people may laugh more than others. I bet several of these are also plausible explanations for the difference in doctor's visits between the groups. That is, there are a whole variety of *additional* differences between people who laugh a lot and those who laugh little, any or all of which could be responsible for the differences in health *in addition to* or *instead of* laughing per se. That is, our laughter "IV" is *inevitably* confounded.

This is true of all such variables. They are inevitably confounded. Thus, we can only make suggestions and guesses about causality. It is only when we are free to allocate people to levels randomly that we have the possibility of breaking the natural associations between variables and thus can remove these confounds from our design.

To understand this, imagine that instead of *measuring* the laughter variable and *classifying* participants into high and low laughter groups, we *manipulate* this variable. There are a number of ways that we could do this, but let's keep it simple. Say that we randomly allocate participants to one of two conditions. Those in the *experimental* condition we get to laugh out loud for 1 full minute every 3 waking hours for 5 days. Those in the *control* condition we get to hum a tune of

their choice for 1 full minute every 3 waking hours for 5 days. We use the same DV as before and get the same results: the laughter group pays fewer visits to the doctors. If the randomization has worked properly, then there should be no systematic differences between the groups in the variables that confounded the other version of this study: personality, stress, happiness, and so on. We would then be in a position to attribute unequivocally any differences in physical health to laughter.

You will come across many examples of such variables in your reading and will use them in your own studies in psychology. It is perfectly appropriate to use such variables in studies and analyses, provided that you are circumspect in the conclusions you reach. You can run studies with such variables in conjunction with true IVs (e.g., a study on the effects of alcohol and gender on driving). However, if you *only* have such variables in your design, make sure that you are aware that it is *not* an experiment. You therefore need to be able to differentiate a true IV from something that looks like one but is not.

? SAQ 48 Look again at the studies in SAQ 45. Are all the variables true IVs? State for each study whether this is the case and which variable(s) are not true IVs.

Because of this, some researchers suggest that we should replace the terms independent and dependent variables when using such variables with ones that do not imply causality. One possibility is to refer instead to the **explanatory** variable and the **response** variable. Other possibilities include **classification** variable, **subject** variable or **predictor** variable instead of IV and **measured** variable or **criterion** variable instead of DV. Your tutor may encourage you to use such alternative labels, so watch out for this issue.

One relevant term that you may also come across in your reading is **quasi-experimental design**. A quasi-experimental design is one that looks like an experimental design but that lacks the critical ingredient – random assignment. Quasi-experimental variables are thus particular versions of these variables that look like IVs but are not really independent variables. Quasi-experimental designs are interesting and useful methods of attempting to conduct experiments in real-life settings where it is not possible to randomly assign participants to treatment conditions. You may learn more about such designs in your methods course.

Summary of Section 13.8

1 Sometimes you will include in your experiments variables that look like IVs but that are not true independent variables.

2 For a variable to be a true independent variable you must be able to randomly assign participants to the levels of your IV or to randomly assign orders of conditions to participants.

3 Quasi-experimental variables are ones in which participants are in different conditions but have not been assigned to these conditions randomly. They are thus a particular class of variables that look like IVs but are not true IVs.

13.9 Some tips to help you to design better experiments and write better reports

This brings us almost to the end. I will spend the final few pages of the book giving you a few tips to help you to design better experiments and, I hope, also to write better reports. You can find some more tips for your later experiments – those that you design as you become more experienced – in Section G of the book's Web site at **http://mcgraw-hill.co.uk/openup/harris/**

13.9.1 The basic rule

A very important rule to remember is that there is no such thing as *the* right way to design an experiment to test a given hypothesis. Although there will be lots of wrong ways, there will usually be several right ways, perhaps many. So, do not feel overawed by the task. Relax! Your task is to design a reasonable, meaningful experiment, free of confounding variables, not to find the "correct" way to design it.

13.9.2 Getting reliable measures of the dependent variable

Would it be fair if your degree classification or grade for any other major qualification were based on a single, end-of-course exam? If I wanted to find out how good you were at something – say, bowling – would it be a good idea to judge you after a single roll? In both cases

the answer is, of course, no. A range of factors, not just ability, influences performance on any task. For any single performance, such as one exam, these factors may serve to make you perform below or above par. This is why you tend to be assessed many times on a course.

Well, exactly the same arguments apply to the performance of your participants in an experiment. A common weakness of many student experiments is to measure the participant's performance once. This is unlikely to be *reliable*. If you can, you should take enough measures from a participant to be reasonably confident that you have a *reliable* estimate of their scores on the DV in each condition of the experiment in which they appear. A **reliable** measure would give us a similar value on the DV for a given participant if we had the luxury of assessing him or her under identical conditions on different occasions. A reliable measure is essential if we are going to be able meaningfully to assess the effects of the IV.

Thus, wherever possible, experiments should have more than one *trial* per condition. A **trial** in this context is a basic unit of experimental time. It begins when you present something to the participant and ends when they give you back a response on your DV.

However, as with all aspects of good experimental design, you need to use your common sense. You should have as many trials (i.e., take as many separate measures) of the DV as you need, but not so many as *to induce fatigue or boredom*. The optimum number of measures is one of the things that you can establish during the piloting phase of the study (see below). You should also watch out for situations in which it is not possible to have more than one trial per condition (e.g., when doing so will ask too much of your participants, creates significant carry-over effects, or might reveal too much to your participant about the purpose of the study).

13.9.3 Pilot testing

Once you have designed your experiment, chosen the question you wish to examine experimentally, the IV you wish to manipulate, the DVs you think will best assess this manipulation, decided whether you would be better off comparing the same or different participants, prepared your materials, set up your random sequences, and standardized your instructions, you might think that you are finally ready for the off. However, a little more patience at this stage may well pay

dividends later. For, rather than diving straight in to running your study, a sensible procedure to adopt here is to **pilot test** your experiment. That is, to try it out on a few participants first to see whether it makes sense to them, to uncover any serious flaws or problems that might have been overlooked at the design stage and to generally "fine-tune" the procedure. Pilot testing also enables you to familiarize yourself with your role as experimenter so that you are practised and professional by the time you encounter your first participant proper.

Pilot testing can save you a lot of wasted time and effort. It provides you with a golden opportunity to improve the design *before* you have wasted too many participants. It is, therefore, a good habit to get into. Indeed, it will become increasingly important as you take greater responsibility for designing your own experiments.

A pilot test can also reveal whether you have a potential *floor* or *ceiling* effect in your data. For instance, thinking back to the driving and music experiment, we need to set our participants a course on the driving simulator that is neither so difficult that few can do it (**floor effect**), nor so easy that more or less everyone can (**ceiling effect**). Otherwise, we will not be able to assess the effects of manipulating our IV on performance itself. However, it is often difficult to decide on the appropriate level of task difficulty in the abstract. A brief pilot test on a small sample of participants who are *representative of the sample that will take part in the actual experiment* can help with this problem. (It is no good piloting on your relatives and their friends but running the experiment on a sample of students.)

Do do it – even one or two participants may tell you something useful about your study. (Of course, you should run more than one or two if you can.)

13.9.4 The post-experimental interview

Once you have finished collecting all the data you need from a participant, it is often a good idea to spend a few minutes with him or her before debriefing (Section 10.10.2) conducting a **post-experimental interview**. Set yourself up a little interview schedule with gently probing questions to try to find out more about what crossed participants' minds while they did your experiment. This may give you insights into factors controlling their behaviour and the extent to which they were aware of the hypotheses of the experiment. It may help you to interpret your findings and even give you ideas for future research.

If possible, someone who is blind to condition should conduct the interview. Such interviews are an essential feature of pilot tests and can also be useful in the main experiment. Unlike the debriefing, the post-experimental interview is primarily for *your* benefit, not the participants'.

13.9.5 Check and screen your data prior to statistical analysis

After putting all the effort and time into designing the experiment, running it, and collecting the data, an odd thing can happen. You can find yourself suddenly not caring about how accurately the data have been entered into the computer for analysis and just keen (desperate?) to get the analysis done and dusted. When you enter the data yourself this can be because you cannot imagine that you have made a mistake (as *if*). However, it happens even when someone else has entered the data for you. I suspect that it is because you are so close to completion and is part of the general need that we have for closure and for starting the next task on our list. It is important to guard against this tendency, of course.

Whenever you collect any data, whether from an experiment or any other study, you *must* ensure that you have entered them accurately prior to analysis. It is easy to make mistakes when doing something as dull as entering data into a computer. **Check the data** entry in at least two ways. First, get on your screen a frequency count, histogram or bar chart of the IV and the DV, just to check that all the values are ones that make sense (i.e., that they all fall in the right range of values). (Print this out only if you can be sure that it will not cost the earth in paper.) For example, if you find a value of 9 or 33 for a 7-point scale or of 222 for age, you *know* that this is outside the range of possible or legal values. It *must* be a mistake. However, this technique cannot help you spot values that are legal but wrong (for example, giving someone a score of 1 instead of 6 on a 7-point scale). So, second, **proofread** the data by checking that the numbers in the data window for a given participant match those on whatever source was used to input the data (e.g., the relevant coding sheet or questionnaire). If you have lots of participants, do this for about 25 percent of the sample in the first instance, making sure that you sample the data entered early, in the middle and late on. If you discover errors, then check more. If you do not have many participants, check the data for them all just to make sure.

Once you are satisfied that the data no longer contain mistakes, it is a good idea to use graphs and other techniques to check whether the data meet the assumptions of the tests that you hope to use. Look in your statistics textbook for ways of displaying data using histograms, bar charts, scatter plots, stem and leaf plots, and box plots. Most statistical software packages will provide ways of producing such displays and may even contain useful programs to help you with this (e.g., EXPLORE in SPSS for Windows). You will learn more on your statistics course about what to do and what to look out for. The latter include extreme values that unduly influence the means and standard deviations (known in the trade as **outliers**) and any *missing data*. If **screening** your data in this way ever leads you to change or transform the data, make sure that you report what you did and why in the opening paragraph of your RESULTS.

Only once you have done these things should you begin the analyses. You must get into the habit of inspecting your data *before* you compute *any* statistics. Otherwise you risk basing your RESULTS on nonsense.

Summary of Section 13.9

1 Where possible, it is a good idea to have more than one trial per condition – that is, to measure your participant's performance more than once in any given condition of the experiment.

2 A good habit to get into is to pilot test your experiment by trying it out on a few participants *before* you start to record any data. This enables you to spot any flaws in your design and gives you the opportunity to fine-tune the procedure. Pilot testing can save you a lot of wasted time and effort.

3 You must, however, run the pilot test on a sample of participants from the same population as the one you will actually use in the experiment.

4 In some experiments a post-experimental interview with participants can give you useful insights into the variables controlling their behaviour in the experiment. This may help you to interpret your findings and can give you ideas for follow-up studies.

5 Make sure that you check that the data have been entered accurately before analysis. As you become more experienced, learn also how to screen your data thoroughly before analysing it.

13.10 Above all, randomize properly

Finally, here's a note to all you budding experimenters out there: *Randomize properly*! The random assignment of participants to conditions or of orders of conditions to participants is the key to the whole experimental enterprise. At all times and in all experiments use *truly* random procedures for assignment or else the exercise is in danger of being pointless. So, how you go about randomization is a central issue (see Appendix 2). You must take this problem seriously and not do it casually. Think carefully in advance about how you will do the randomization. Use random number generators or tables. Do *not* make the numbers up yourself or use a procedure that can be influenced consciously or even unconsciously. Learn to differentiate randomization *with* replacement from randomization *without* replacement and use the procedure that you require. Subsequently, describe in sufficient detail in your reports the way in which you randomized so that your readers can judge for themselves whether the randomization was effective.

That's it! Good luck. I hope you enjoy finding things out with and about experiments in psychology.

Consolidating your learning

This closing section is designed to help you check your understanding and determine what you might need to study further. You can use the list of methodological and statistical concepts to check which ones you are happy with and which you may need to double-check. You can use this list to find the relevant terms in the chapter and also in other textbooks of design or statistics. The diagnostic questions provide a further test of your understanding. Finally, there is a description of what is covered in the two statistics textbooks paired with this book plus a list of relevant further material that you can find on the book's Web site.

Methodological and statistical concepts covered in this chapter

Ceiling effect
Checking and screening data
Explanatory variable (response, classification, subject, predictor variable)
Factorial design
Floor effect
Interaction effect
Line graphs
Main effect
Measured variable (criterion variable)
Mixed design (split plot design)
Outlier
Pilot test
Post-experimental interview
Quasi-experimental design
Reliability
Repeated measures design
Trial
Two by two (2×2) design

Diagnostic questions – check your understanding

1 If we extend the number of levels on an unrelated samples IV are we more likely to face problems with recruiting adequate numbers of participants or with order effects?
2 How many IVs are there in a 2×2 mixed design and what type of IVs are they?
3 What is a significant *main* effect?
4 What does it mean to say that two variables *interact*?
5 When is an unrelated samples IV a "true" IV?

You can find the answers at the rear of the book (p. 278).

Statistics textbooks

∑ The books paired with *Designing and Reporting Experiments* have the following coverage:

Learning to use statistical tests in psychology – Greene and D'Oliveira
Greene and D'Oliveira describe how to analyse data from a study with more than 2 levels on an unrelated samples IV using a non-parametric test in Chapter 12 (Kruskal-Wallis test) and a parametric test in Chapter 14 (one-way ANOVA for unrelated samples). They describe how to analyse data from a study with more than 2 levels on a related samples IV using a non-parametric test in Chapter 11 (Friedman test) and a parametric test in Chapter 15 (one-way ANOVA for related samples). They describe how to use ANOVA to analyse data from studies with two IVs, including mixed designs, in Chapters 17–19. These chapters include discussion of how to interpret and graph interaction effects.

SPSS survival manual – Pallant
Pallant describes how to use SPSS to analyse data from a study with more than 2 levels on an unrelated samples IV using a non-parametric test in Chapter 16 (Kruskal-Wallis test) and a parametric test in Chapter 18 (one-way between-groups ANOVA). She describes how to analyse data from a study with more than 2 levels on a related samples IV using a non-parametric test in Chapter 16 (Friedman test) and a parametric test in Chapter 15 (one-way repeated measures ANOVA). She describes how to use ANOVA to analyse data from studies with two IVs, including mixed designs, in Chapters 19 and 20. She has good coverage of interaction and main effects. Pallant describes how to screen and clean data in Chapter 5 and how to check it for violations of statistical assumptions in Chapter 6. She also explains how to use graphs to describe and explore data (Chapter 7) and has coverage of issues to do with reliability and validity in Chapters 1, 8 and 9.

The *Designing and Reporting Experiments* Web site

The book's Web site at **http://mcgraw-hill.co.uk/openup/harris/** has a number of sections that expand on the issues discussed in this

chapter. Section C contains further consideration of main effects and interactions, including advice on how to graph three-way interactions and a list of possible main effects and interactions to be found in designs with four IVs. Section F contains advice on things to watch out for with IVs that are not true IVs. Section G contains seven tips for advanced students to further improve their experiments. Section I contains pointers to how to use the Web site for specific advice for those designing and reporting final year projects.