Research methods: Design of investigations

- **Aims and hypotheses**
  Why studies are carried out and what they are designed to test.

- **Selecting participants**
  Different ways of selecting the people to take part in a study.

- **Good practice in experimentation**
  Using standardised procedures and avoiding unwanted variables.

- **Experimental designs**
  The three types of design available to experimenters.

- **Good practice in non-experimental designs**
  How to design interview, survey, and case studies, and some pitfalls to avoid.

- **Problems with experimental research**
  Experimenters can influence the outcome of their studies in unexpected ways.

- **General issues in investigations**
  Other things that can and sometimes do affect the results of psychological studies.

Independent and dependent variables
One- and two-tailed hypotheses
Null hypothesis
Manipulating and measuring the variables

Random and systematic sampling
Stratified and quota sampling
Opportunity sampling
Sample size

Standardised procedures
Confounding and controlled variables
Random error
Operationalisation

Independent design
Matched participants design
Repeated measures design and counterbalancing

Evaluation apprehension and demand characteristics
Participant and non-participant observation
Survey, questionnaire, and attitude scale design
Correlational studies

Experimenter’s effects
Rosenthal’s flatworm study
Barber’s list of nine effects
Demand characteristics and evaluation comprehension

The Hawthorne effect
Validity and generalisability
Meta-analyses
Reliability and replication

© 2004 Psychology Press Ltd
In order to carry out a study successfully, care and attention must be devoted to each stage in its design and implementation. This chapter is concerned with these issues mostly with respect to experimental designs. However, there is also full consideration of the factors involved in producing good non-experimental designs.

As we will see, several decisions need to be made when designing an experimental study:

1. The investigator must decide what he or she hopes to achieve by carrying out the study. This involves generating appropriate aims and hypotheses.
2. The investigator has to work out how the variables specified in the hypotheses are to be manipulated and/or measured.
3. Appropriate procedures need to be used when selecting participants for the study.
4. Attention needs to be paid in the experimental design to ensuring that the effects of any situational variables on the participants’ behaviour are minimised.
5. If the investigator is using an experimental design, then he or she has to select an appropriate one. This includes a decision as to whether each participant will be included in only one condition or in both conditions.
6. Care has to be paid to the relationship between the participants and the investigator in order to prevent systematic biases in the data obtained.

The success or otherwise of the investigator’s study can be evaluated in terms of various criteria. If the design and its implementation are appropriate, then the reliability of the findings and their replicability will tend to be high. In addition, use of an appropriate design maximises the validity of the findings.

**AIMS AND HYPOTHESES**

The first step that needs to be taken when designing an experimental or non-experimental study is to decide on the aims and hypotheses of the study. The aims are usually more general than the hypotheses, and they help to explain the reasons for the investigator deciding to test some specific hypothesis or hypotheses. In other words, the aims tell us *why* a given study is being carried out, whereas the hypotheses tell us *what* the study is designed to test.

**Experimental studies**

The distinction between aims and hypotheses can be seen more clearly if we consider an example. Suppose that we decide to test the levels-of-processing theory put forward by Craik and Lockhart (1972), which states that information that has been processed for meaning will be remembered better than information that has not. In order to do this, we might present all of our participants with the same list of nouns and then ask them to provide free recall 30 minutes later. Half of them might be asked to think of adjectives to go with the nouns (processing of meaning or semantic processing), whereas the other half are asked to think of rhyming words (non-semantic processing). In such a study, the main aim is to investigate levels-of-processing theory. In more general terms, the aim is to see whether long-term memory is influenced by the kind of processing that occurs at the time of learning. The experimental hypothesis is more specific: free recall from long-term memory is higher when there is semantic processing at the time of learning than when there is non-semantic processing.

**Non-experimental studies**

The situation with regard to the aims and hypotheses is somewhat different in qualitative research, in which the data are not in numerical form. Qualitative research is often based on interviews, observations, or case studies. Qualitative researchers frequently have no
specific hypotheses at the outset of the study; rather, the hypotheses to be tested emerge from a detailed consideration of the data. The aims of qualitative research tend to be more general and wide-ranging than those of traditional research.

An example of qualitative research is the work of Marsh, Rosser, and Harré (1978) with football fans. Marsh’s original aim was to try to understand the aggressive behaviour that they often display, but he had few if any preconceptions or hypotheses in mind at the outset of the study. During the course of the study, Marsh et al. (1978) began to realise that there were complex rules or social norms that were shared by football fans, and which played an important role in determining their behaviour (see PIP, Chapter 20).

Hypotheses

Most experimental research starts with someone thinking of an experimental hypothesis (also known as the alternative hypothesis). This is simply a prediction or expectation of what will happen in a given situation. For example, you might think of the experimental hypothesis that loud noise will have an effect on people’s ability to carry out a task, such as learning the information in a chapter of an introductory psychology textbook.

Variables

As with most experimental hypotheses, the one just mentioned predicts that some aspect of the situation (in this case, the presence of loud noise) will have an effect on the participants’ behaviour (in this case, their learning of the information in the chapter). In more technical language, the experimental hypothesis refers to an independent variable, which is usually some aspect of the experimental situation that is manipulated by the experimenter. In our example, the presence versus absence of loud noise is the independent variable. The hypothesis also refers to a dependent variable, which is some aspect of the participants’ behaviour. In our example, some measure of learning would be used to assess the dependent variable. In a nutshell, most experimental hypotheses predict that a given independent variable will have some specified effect on a given dependent variable.

One-tailed or two-tailed?

It should be noted at this point that there are two types of experimental hypothesis: directional or one-tailed hypotheses, and non-directional or two-tailed hypotheses. A directional or one-tailed hypothesis predicts the nature of the effect of the independent variable on the dependent variable. In terms of our example, a directional hypothesis might be as follows: loud noise will reduce people’s ability to learn the information contained in the chapter of a textbook. A non-directional or two-tailed hypothesis predicts that the independent variable will have an effect on the dependent variable, but the direction of the effect is not specified. In terms of our example, a non-directional hypothesis would be as follows: loud noise will have an effect on people’s ability to learn the information contained in the chapter of a textbook. This hypothesis allows for the possibility that loud noise might actually improve learning.

Null hypothesis

The experimental hypothesis consists of the predicted effect of the independent variable on the dependent variable. This can be contrasted with the null hypothesis. The null hypothesis simply states that the independent variable will have no effect on the dependent variable. In terms of our example, a suitable null hypothesis would be as follows: loud noise will have no effect on people’s ability to learn the information contained in the chapter of a textbook.
to learn the information contained in the chapter of the textbook. In a sense, the purpose of most studies using the experimental method is to decide between the merits of the experimental hypothesis and those of the null hypothesis.

Why do we need a null hypothesis when what we are interested in is the experimental hypothesis? The key reasons are because the null hypothesis is much more precise than the experimental hypothesis, and we need precise hypotheses in order to use statistical tests properly. For example, the null hypothesis that loud noise will have no effect on people’s learning ability is precise because it leads to the prediction that the single most likely outcome is that performance will be equal in the loud noise and no noise conditions. Failing that, there will probably only be a small difference between the two conditions, with the difference being equally likely to go in either direction. In contrast, consider the experimental hypothesis that loud noise will reduce people’s learning ability. This hypothesis is very imprecise, because it does not indicate how much learning ability will be impaired. This lack of precision makes it impossible to decide the exact extent to which the findings support or fail to support the experimental hypothesis.

**Manipulating the independent variable**

It might seem easy to do a study to test the experimental hypothesis that loud noise disrupts learning. However, there are various pitfalls that need to be avoided. The first issue that needs to be considered is how to manipulate the independent variable. In our example, we want to compare loud noise with no noise, so we have to decide exactly how loud we want the noise to be. If it is very loud, then it might damage the hearing of our participants, and so would be totally unacceptable. If it is fairly soft, then it is unlikely to have any effect on the learning ability of our participants. It is also likely to make a difference whether the noise is meaningful (e.g. music or speech) or meaningless (e.g. the noise of a road drill).

**Measuring the dependent variable**

The second issue is how to measure the dependent variable or aspect of the participants’ behaviour. We could ask the participants various questions to measure their understanding of the material in the textbook chapter. However, selecting the questions so that they are not too easy or too hard requires careful thought.

**SELECTING PARTICIPANTS**

Studies in psychology rarely involve more than about 100 participants. However, researchers generally want their findings to apply to a much larger group of people than those acting as participants. In technical terms, the participants selected for a study form a **sample**. This sample is taken from some larger **population**, which consists of all the members of the group from which the sample has been drawn. For example, we might select a sample of 20 children aged 5, for a study, in which case the population might consist of all the 5-year-olds living in England.

When we carry out a study, we want the findings obtained from our sample to be true of the population from which they were drawn. In order to achieve this, we must use a **representative sample**, i.e. participants who are representative or typical of the population in question. However, numerous studies have been carried out with
non-representative samples; the term *sampling bias* is used to refer to this state of affairs. Coolican (1994, p. 36) was pessimistic about the chances of selecting a representative sample:

*The simple truth is that a truly representative sample is an abstract ideal unachievable in practice. The practical goal we can set ourselves is to remove as much sampling bias as possible.*

**Random samples**

To return to our earlier example, we might study the effects of loud noise on learning in students preparing for a psychology exam. The best way of obtaining a representative sample from that population would be to make use of *random sampling*. We could obtain lists of names of all the students due to sit the psychology exam in a given year. After that we could use some random method to select our sample. This could be done by picking names out of a hat, or by sticking a pin repeatedly into the lists.

Another approach is to assign a number to everyone in the population from which the sample is to be selected. After that, a computer can be used to generate a series of random numbers that can be used to select the sample. Alternatively, random number tables can be used in a similar way to produce the sample.

If we wanted to have a representative sample of the entire adult population, then we could apply one of the methods of random selection just described to the electoral roll. However, even that would be an imperfect procedure. Several groups of people, including the homeless, illegal immigrants, and prisoners, are not included in the electoral roll.

As Cardwell et al. (1996) pointed out, there is a modified version of random sampling which is easier to use. This is *systematic sampling*. It involves selecting the participants by a quasi-random procedure. For example, if we have a list of all the members of the population, we could select every hundredth name from that list as participants. This procedure is not as effective as random sampling because it cannot be claimed that every member of the population is equally likely to be selected.

Random sampling typically fails to produce a truly representative sample, because it is actually very hard for an experimenter to obtain a random sample. There are various reasons for this. First, it may not be possible to identify all of the members of the larger population from which the sample is to be selected. Second, it may not be possible to contact all those who have been selected randomly to appear in the sample.

Third, some of those who are selected to be in the sample are likely to refuse to take part in the study. This might not matter if those who agreed to take part in research were very similar in every way to those who did not. However, there is considerable evidence that volunteers differ in various ways from non-volunteers. Manstead and Semin (1996, p. 93) discussed some of the evidence, and concluded, “there are systematic personality differences between volunteers and non-volunteers.” Volunteers tend to be more sensitive to the demand characteristics (cues used by participants to work out what a study is about), and they are also more likely to comply with those demand characteristics.

In sum, it is worth bearing in mind what Coolican (1998, p. 720) had to say about random samples: “Many students write that their sample was ‘randomly selected’. In fact, research samples are very rarely selected at random.”

**Stratified and quota samples**

Another way of obtaining a representative sample is by using what is known as *stratified sampling*. The first step is to decide which characteristics of the population might be relevant for the study we want to carry out. These characteristics might include gender and the part of the country in which people live. The second step is to select a representative sample from each stratum.

Random sampling is not as effective as systematic sampling, but it does help to overcome the biases of the researcher. If we select every hundredth name on the list, we avoid missing out names that we cannot pronounce, or do not like the look of.

Why do you think volunteers are more likely than non-volunteers to be sensitive to the design characteristics of a study?

**Key Terms**

- **Random sampling**: selecting participants on some random basis (e.g. coin tossing).
- **Systematic sampling**: a modified version of random sampling in which the participants are selected in a quasi-random way (e.g. every hundredth name from a population list).

Stratified sampling is not as effective as random sampling, but it does help to overcome the biases of the researcher. If we select every hundredth name on the list, we avoid missing out names that we cannot pronounce, or do not like the look of.

**Systematic sampling**: a modified version of random sampling in which the participants are selected in a quasi-random way (e.g. every hundredth name from a population list).

- **Stratified sampling**: a modified version of quota sampling, in which the selection of participants according to certain characteristics is decided by the researcher, rather than in a random way.
which they live. This allows us to think in terms of sub-groups. After that, we select participants at random from within each of the sub-groups.

Suppose that we want to carry out a study on A-level psychology students. We know that 75% of A-level psychology students are female, and that 40% of all A-level psychology students live in the north of England. We could then ensure that the participants used in our experiment were selected in a random way so that 75% of them were female, and 40% of them lived in the north of England. If we make use of enough criteria, then stratified sampling can be an effective way of finding a representative sample.

There is a modified version of stratified sampling which is known as quota sampling. It resembles stratified sampling in that participants are selected in proportion to their representation in the population. However, it differs in that the researcher decides who to include in each sub-group, rather than the decision being made at random. Quota sampling is often used in market research. It tends to be faster than stratified sampling. However, it has the disadvantage that people who are readily available (e.g. the unemployed) are more likely to be included than those who are not.

The problem with stratified and quota sampling is that it is often hard to know which sub-groups to identify. It is a waste of time and effort if we use characteristics (e.g. gender) that are of no relevance to the study. What is more troublesome is if we fail to identify sub-groups on the basis of some characteristic (e.g. GCSE performance) which is actually highly relevant.

**Opportunity sampling**

Random sampling, stratified sampling, and quota sampling are often expensive and time-consuming. As a result, many researchers use opportunity sampling. This involves selecting participants on the basis of their availability rather than by any other method. Opportunity sampling is often used by students carrying out experiments, and it is also very common in natural experiments (see the Research methods: Psychological enquiry chapter).

Opportunity sampling is the easiest way to proceed. However, it has the severe disadvantage that the participants may be nothing like a representative sample. For example, students who are friends of the student carrying out a study may be more willing to take part than students who are not.

**Sample size**

One of the issues that anyone carrying out a piece of research has to consider is the total number of participants to be included. What is the ideal number of participants in each condition? There is no definite answer to that question, but here are some of the relevant factors:

- It is generally expensive and time-consuming to make use of large samples running into hundreds of participants.
- If it requires very large samples to obtain a statistically significant effect of some independent variable on some dependent variable, then this suggests that the effect is small and of little practical importance.
- If we use very small samples (fewer than 10 participants in each condition), then this reduces the chances of obtaining a significant effect.
- In general terms, sampling bias is likely to be greater with small samples than with large ones.

If there is a golden rule that applies to deciding on sample size, it is the following:

*The smaller the likely effect being studied, the larger the sample size needed to demonstrate it.*

For most purposes, however, having about 15 participants in each condition is a reasonable number.
GOOD PRACTICE IN EXPERIMENTATION

In order for an experiment to be designed and carried out successfully, there are several considerations that the researcher needs to bear in mind. Some of the main considerations are discussed in detail in this section.

Standardised procedures

In order to carry out an experiment successfully, it is very important that every participant in a given condition is treated in the same way. In other words, it is necessary to use standardised procedures. For example, consider the instructions that are given to the participants. In order to ensure that all of the participants get precisely the same instructions, the experimenter should write them down. He or she should then either read them to the participants, or ask the participants to read them to themselves.

In similar fashion, standardised procedures should be used for the collection of data. Suppose we want to assess the effects of loud noise on learning from a book chapter. We might ask the participants to write down everything they could remember about the chapter. However, it would be very hard to compare the recalls of different participants with any precision. A standardised procedure would be to ask all of the participants the same set of, say, 20 questions relating to the chapter. Each participant then obtains a score between 0 and 20 as a measure of what he or she has learned.

Is it easy to make sure that standardised procedures are being used? No, it is not. Most experiments can be thought of as social encounters between the experimenter and the participant, and it is customary to behave in different ways towards different people. Robert Rosenthal (1966) studied some of the ways in which experimenters fall short of standardised procedures. He found, for example, that male experimenters were more pleasant, friendly, honest, encouraging, and relaxed when their participants were female than when they were male. This led him to conclude as follows: “Male and female subjects [participants] may, psychologically, simply not be in the same experiment at all.”

Confounding variables

Another issue to consider is whether or not our experiment contains any confounding variables. These are variables that are mistakenly manipulated along with the independent variable. Suppose there is a study in which one group of participants receives no noise and reads a chapter at midday, whereas the other group of participants receives loud noise and reads the same chapter at midnight. If we find that the latter group learns less well than the former group, we would not know whether this was because of the loud noise or because they did their learning late at night when they were very tired. In this example, time of day is a confounding variable.

Confounding variables are especially likely to be found in non-experimental investigations in which the researcher has no control over the independent variable. One of the classic examples concerns the work on maternal deprivation that was carried out on institutionalised children (see PIP, Chapter 17). Bowlby (1951) argued that these children had poorer social and intellectual development than other children because of the absence of the mother. However, these children also had to cope with the unstimulating environment of the institutions of those days, and this was a confounding variable that Bowlby tended to ignore.

Confounding variables are a form of constant error. Constant error is present when the effects of any unwanted variable on the dependent variable differ between conditions. There are numerous types of constant error. The participants in one condition may be
more tired than those in another condition, or they may be more intelligent, or they may be more motivated.

**Controlled variables**

How do we avoid having any confounding variables? One useful approach is to turn them into **controlled variables**, which are variables that are held constant or controlled. Suppose that we want to study the effects of noise on learning, and we are concerned that the time of day may have an effect. We could make time of day into a controlled variable by testing all of our participants at a given time of day, such as late morning or early evening. If we did this, we would know that time of day could not distort our findings.

**Random error**

**Random error** occurs when variables that are totally irrelevant to the experiment influence the behaviour of the participants. The key difference between random error and constant error is that random error generally affects both conditions equally and so has **unsystematic effects**, whereas constant error has **systematic effects** on one condition but not on the other. Constant error is more serious than random error, because it can lead us to misinterpret our findings. However, random error is also of concern, because it introduces unwanted variation in the dependent variable.

There are almost limitless types of random error. For example, suppose we are interested in comparing learning performance under noise and no-noise conditions. Participants in either condition may learn poorly because they have a splitting headache, because they have just argued with a close friend, because a relationship broke up last week, because the weather is bad, or because they are worried about an important examination they have to take next week. The experimenter cannot control most forms of random error, but should try to control those that can be controlled. For example, he or she should ensure that the lighting conditions, the heating conditions, the experimenter’s tone of voice, and so on remain constant for all participants.

**Operationalisation**

Psychologists carry out studies to test experimental hypotheses, such as “anxiety impairs performance” or “maternal deprivation leads to maladjustment”. There is an immediate problem with designing a study to test such hypotheses: there is little or no agreement on the best way to measure psychological concepts or variables such as “anxiety”, “performance”, “maternal deprivation”, or “maladjustment”. The most common approach to this problem is to make use of **operationalisation**. This involves defining each variable of interest in terms of the operations taken to measure it. Such a definition is referred to as an operational definition. For example, anxiety might be defined in terms of the score on a self-report questionnaire largely ignores physiological and behavioural aspects of anxiety, and no-one believes that performance can only be assessed in terms of rate of anagram solution.
In spite of these important limitations with operational definitions, it is hard to carry out research without using them. Stretch (1994, p. 1076) argued that operational definitions should be used in a careful fashion:

*A useful rule of thumb is to consider many different ways of measuring the psychological construct of interest and determine the extent to which each method could yield different experimental results. If you find that the measurement techniques radically affect the results that emerge, this should indicate that more work is needed on developing the underlying psychological and measurement models to explain these effects.*

**EXPERIMENTAL DESIGNS**

If we wish to compare two groups with respect to a given independent variable, it is essential to make sure that the two groups do not differ in any other important way. This general rule is important when it comes to selecting participants to take part in an experiment. Suppose all the least able participants received the loud noise, and all the most able participants received no noise. We would not know whether it was the loud noise or the low ability level of the participants causing poor learning performance.

How should we select our participants so as to avoid this problem? There are three main types of experimental design:

- **Independent design**: each participant is selected for only one group.
- **Matched participants design**: each participant is selected for only one group, but the participants in the two groups are matched for some relevant factor or factors (e.g. ability; sex; age).
- **Repeated measures design**: each participant appears in both groups, so that there are exactly the same participants in each group.

**Independent design**

With the independent design, the most common way of deciding which participants go into which group is by means of randomisation. In our example, this could involve using a random process such as coin tossing to decide whether each participant is exposed to loud noise or to no noise. It is possible with randomisation for all the most able participants to be selected for the same group. However, what happens in the great majority of cases is that the participants in the two groups are similar in ability, age, and so on.

**Matched participants design**

With the matched participants design, we make use of information about the participants to decide which group each participant should join. In our example, we might have information about the participants’ ability levels. We could then use this information to make sure that the two groups were matched in terms of range of ability.

**Repeated measures design**

With the repeated measures design, every participant is in both groups. In our example, that would mean that each participant learns the chapter in loud noise and that they also learn the chapter in no noise. The great advantage of the repeated measures design is that we do not need to worry about the participants in one group being cleverer than those in the other group: as the same participants appear in both groups, the ability level (and all other individual characteristics) must be identical in the two groups!

The main problem with the repeated measures design is that there may well be order effects. Their experiences during the experiment may change the participants in various ways. They may perform better when they appear in the second group because they have
gained useful information about the experiment or about the task. On the other hand, they may perform less well on the second occasion because of tiredness or boredom. It would be hard to use a repeated measures design in our example: participants are almost certain to show better learning of the chapter the second time they read it, regardless of whether they are exposed to loud noise.

**Counterbalancing**

Suppose we used a repeated measures design in which all of the participants first learned the chapter in loud noise and then learned it in no noise. We would expect the participants to show better learning in no noise simply because of order effects. A better procedure would be to have half the participants learn the chapter first in loud noise and then in no noise, while the other half learn the chapter first in no noise and then in loud noise. In that way, any order effects would be balanced out. This approach is known as **counterbalancing**. It is the best way of preventing order effects from disrupting the findings from an experiment.

**GOOD PRACTICE IN NON-EXPERIMENTAL DESIGNS**

There are several kinds of non-experimental studies (see the Research methods: Psychological enquiry chapter). They include naturalistic observation, participant observation, studies based on correlational analysis, interviews and surveys, and case studies. Case studies involve the collection of detailed information from individuals rather than from groups of participants. We will begin by considering some general points that need to be taken into account when designing and implementing a non-experimental study.

**General considerations**

One of the key issues in many non-experimental studies is to decide whether the participants should be made aware of the fact that they are taking part in research. The main
argument for making the participants aware of what is happening is an ethical one. Voluntary informed consent is regarded as of central importance in ensuring that research is ethically acceptable, and it is impossible to obtain that consent from people who do not know they are taking part in a study. However, participants who are made aware may fail to behave in a natural way. Their behaviour may be affected by evaluation apprehension (desire to impress the experimenter) or by their guesses about the experimental hypothesis being tested (demand characteristics).

Some of the issues to be considered are the same as those that apply to experimental studies. For example, if the participants are intended to form a representative sample from a larger population, then it is desirable to use a suitable form of sampling (e.g. random sampling or systematic sampling). The issue of sampling is perhaps especially important with respect to case studies, which involve intensive investigation of individuals. However, there are many non-experimental studies in which the investigator has little or no control over the selection of participants.

Observational studies
Observational studies differ from each other in various ways. First, we can distinguish between participant observation research, in which the investigator is involved in the study as an active participant, and non-participant observation research, in which the investigator only observes the behaviour of the participants.

Second, there is a distinction between unstructured observation and structured observation. According to Dyer (1995, p. 153), unstructured observation is research “where the aim is simply to ensure that everything which appears to be of relevance to the research at a given moment is recorded.” In contrast, an investigator using structured observation makes prior decisions about what to observe, and this “renders the research process relatively inflexible and incapable of responding to unpredictable situations” (1995, p. 154).

Participant observation
The key factor in participant observation is that the researcher has to do his or her best to become accepted by the social group being studied. The goal is to develop a good understanding of what it is like to be a member of that group, and this can only be done when its members accept and trust the researcher. It follows that participant observation research is very time-consuming, because it can take weeks or months for the researcher to gain the confidence of group members.

Dyer (1995) discussed three stages that are involved in carrying out a participant observation study:

1. Entering the field: an important early step is to be accepted by the “gatekeeper” who controls access to the group to be studied; in a school, it is likely to be the headteacher. It is usually desirable to let the fact that you are doing research emerge gradually over time. However, there are major ethical issues to be considered, and it is important to have the informed consent of those responsible for the running of the school or other organisation.

2. Being in the field: for the duration of the study, you have the hard task of trying to fit in as a member of the group and of remaining detached as an observer. You should take field notes, which are an extensive record of what members say and do. These field notes should be condensed into a field diary that is written up every day, and which should identify key themes. Finally, the field diary is used as the basis for the research report. The initial field notes might include information in the following categories suggested by Lovland (1976): acts (short actions); activities (actions taking up at least several days); meanings (participants’ explanations for their actions); participation (the various roles participants play); relationships among the group members; and settings (the situations in which the group members find themselves).

KEY TERMS
Evaluation apprehension: the desire of participants in a study to be evaluated positively by the experimenter.
Demand characteristics: cues used by participants to try to work out what the experiment is about.
3. Leaving the field: there are major ethical issues in participant observation research, because it tends to deal with personal and sensitive issues. It is thus very important to make sure that the group members have the chance to read and comment on your research, and that you take their comments seriously.

Non-participant observation

Most non-participant observation research starts with the researcher thinking of an experimental hypothesis. If structured observations are to be made, it is then necessary to devise the behavioural categories that are going to be used by the observers. The categories should possess the following features:

1. They should be defined in a reasonably precise and objective way so there is as little ambiguity as possible about which forms of behaviour by the participants qualify for each category.
2. The system of categories needs to be comprehensive, in the sense that all aspects of behaviour that are relevant to the experimental hypothesis should be included.
3. The categories should be usable in the context of the study. For example, a researcher studying the reactions of drivers stuck in traffic jams might include various categories of facial expression. This is only sensible provided that the observer is going to be able to see drivers’ facial expressions clearly from his or her viewing position.

Another key decision concerns the way in which the participants’ behaviour is to be sampled. Dyer (1995) identified four possible sampling procedures:

1. Continuous observation: the observer records behaviour falling into the various categories non-stop over a fairly lengthy period of time (e.g. 60 minutes).
2. Time-interval sampling: the sampling period is divided into a series of short time intervals (e.g. 60 seconds), and the observer decides whether any given participant produces each category of behaviour during each period. Any behaviour is simply recorded as absent or present, so that no distinction is drawn between a given behaviour exhibited once versus more than once during a single time interval.
3. Time-point sampling: the sampling period is divided into a series of short time intervals, and the observer decides whether the various categories of behaviour are present at the end of each sampling period.
4. Random sampling: this is like time-point sampling, except that the points in time at which behaviour is sampled are selected at random.

Survey studies

The survey method involves collecting information from a large group of individuals. This information is often gathered using questionnaires, but can include interviews or phone contacts. It is important in any survey study to ensure that the sample selected is as representative as possible (see earlier). A problem that applies to nearly all sampling methods is that of non-responding. Some individuals who are selected to form part of the sample are likely to refuse to participate. Others agree to participate, but then fail to provide all of the information requested. Persuasion and persistence should be used to minimise the problem of non-responding, but it is very rare for the response rate to be 100% in a survey study.

Survey designs

According to Dyer (1995), there are four main types of survey: one-shot survey; before–after design; two-groups controlled comparison design; and the two-groups before–after design.
One-shot surveys. The one-shot survey is the simplest, but also generally the least informative type of survey. Information is obtained from a single sample at a given point in time. The reason why it is fairly uninformative is that we cannot compare the findings from our sample against those of other groups. As a result, we can only describe what we have found to be the case in the sample we tested.

Before–after surveys. The before–after design is an advance on the one-shot survey, in that data are collected from a single sample on two occasions. The design is most likely to produce interesting findings if some major event or experience intervenes between the first and second data collections. For example, attitudes in the UK towards the Labour party could have been obtained shortly before and after the general election of 1997. Suppose (as seems to have been the case) that attitudes were more positive after the election than before. This might have been due to the election victory, but there are other possibilities. Attitudes to the Labour party might have become more positive even if there had not been an election, or it may be that people tend to respond differently on the second occasion that an attitude questionnaire is completed than on the first. In general, it is hard to interpret the findings based on the before–after design.

Two-groups controlled comparison surveys. The two-groups controlled comparison design is potentially more informative than the designs discussed so far. In essence, there are two similar groups of participants, one of which is exposed to some treatment before data collection, whereas the other is not. For example, attitudes towards the opposite sex could be assessed in those who have (or have not) recently experienced the breakdown of a heterosexual relationship. If the former group was more negative in their attitudes, it could be argued that this was due to the breakdown of the relationship. However, this requires the assumption that the two groups had the same attitudes before the breakdown occurred, and we cannot be sure that that assumption is justified.

Two-groups before–after surveys. The two-groups before–after design is an advance on the two-groups controlled comparison design. Two samples or groups are tested for the first time, then one group is exposed to some treatment before data collection, and finally both groups are tested for a second time. Dyer (1995) gave as an imaginary example a study in which the participants are allocated at random to two groups. The attitudes of all of them towards Third World issues are assessed. One group is then exposed to several presentations of a television commercial focusing on the need to provide economic aid to Third World countries. Finally, the attitudes of both groups towards Third World countries are assessed. This survey method is the most complicated one to use, but the findings are easier to interpret than those from other survey methods.

Questionnaire construction

In order to address the specific issues that interest them, researchers using the survey method often construct their own questionnaire. The first step is to generate as many ideas as possible that might be relevant to the questionnaire. Then those ideas that seem of little relevance are discarded, working on the basis (Dyer, 1995, p. 114) that:

*It is better to ask carefully designed and quite detailed questions about a few precisely defined issues than the same number on a very wide range of topics.*

Closed and open questions. There is an important distinction between closed and open questions. Closed questions invite the respondent to select from various possible answers (e.g. yes or no; yes, unsure, or no; putting different answers in rank order), whereas open questions allow respondents to answer in whatever way they prefer. Most questionnaires use closed questions, because the answers are easy to score and to analyse. Open questions
have the disadvantage of being much harder to analyse, but they can be more informative
than closed questions.

**Ambiguity and bias.** Questions that are ambiguous or are likely to be interpreted in
various ways should be avoided. Questions that are very long or complicated should also
be avoided, because they are likely to be misunderstood. Finally, questions that are biased
should be avoided. Here is an example of a biased question: “In view of the superiority of
Britain, why should we consider further political integration with the rest of Europe?”.

**Reliability and validity.** Good questionnaires need to have high reliability or consistency
of measurement. They also need reasonable validity, meaning that they should measure
what they claim to measure. Reliability can be assessed by means of the test–retest
method, in which the questionnaire is given to the same individuals on two different
occasions. The scores can then be correlated by means of a test such as Spearman’s rho
(see the Research methods: Data analysis chapter). If the correlation is fairly high (about
+0.7 or +0.8), then the questionnaire can be regarded as reliable.

There are several ways of assessing the validity of a test. For example, there is empirical
validity, in which the scores on a questionnaire are compared against some external crite-
 rijon. For example, suppose that someone devised a questionnaire to measure conscien-
tiousness. It seems reasonable to assume that conscientious people will perform better on
examinations, and so we could use examination performance as the external criterion.
Conscientiousness scores on the questionnaire could be correlated with examination
performance using Spearman’s rho, with the assumption being that there would be a sig-
nificant positive correlation.

**Attitude scale construction**

Many of the points made about questionnaire construction also apply to the construction
of attitude scales. However, there are some differences.

**Likert scales.** One of the most common ways to construct an attitude scale is to use the
Likert procedure. Initially various statements are collected together, and the participants’
task is to indicate their level of agreement on a five-point scale running from “strongly dis-
agree” at one end to “strongly agree” at the other end. For positive statements (e.g. “Most
Hollywood stars are outstanding actors”), strongly disagree is scored as 1 and strongly
agree as 5, with intermediate points being scored 2, 3, or 4. For negative statements
(e.g. “Most Hollywood stars are not outstanding actors”), the scoring is reversed so that
strongly disagree is scored as 5 and strongly agree as 1.

Most attitude scales based on the Likert method contain some unsatisfactory items.
One way of finding out which items are unsuitable is by correlating each item separately
with the total score on the scale. Only items that correlate positively at a moderate level
with the total score (+0.3 and above) are retained for the scale.

**Reliability and validity.** The reliability of an attitude scale can be assessed by the
test–retest method. Its validity can generally be assessed by some measure of empirical
validity. For example, we could obtain evidence about the validity of a scale concerned
with attitudes towards religion by correlating the scores with a measure such as regular-
ity of attendance at church, by using Spearman’s rho. However, it is important to note
that the correlation may be low either because an attitude scale lacks validity or because
there is often a large difference between people’s attitudes and their behaviour (see PIP,
Chapter 20).

**Correlational studies**

Correlational studies typically involve obtaining two different measures from a group of
participants, and then assessing the degree of association between the measures by using a
test of correlation such as Spearman’s rho. For example, participants’ level of extraversion
could be correlated with their number of friends, based on the prediction that extraverts are likely to have more friends than introverts.

Correlational studies are easy to carry out. For example, there are thousands of questionnaires for measuring personality or attitudes, and it is possible to take any two at random and administer them to a large group of people. After that, the scores on the two questionnaires can be correlated. However, the fact that correlational studies are easy to perform does not mean that good correlational studies are easily carried out. What features characterise good correlational studies?

**An underlying theory**

First, the study should be based on some underlying theory. The two variables that are measured in the study should both be of clear relevance to the theory. In addition, the predicted direction of the correlation (positive or negative) should follow from the theory. For example, there is the matching hypothesis, according to which we are attracted to those who are about as physically attractive as we are. This was tested in a correlational study by Murstein (1972). The physical attractiveness of couples was judged from photographs. There was a strong positive correlation, with the most physically attractive individuals tending to have someone very attractive as their partner, whereas those who were physically unattractive had unattractive partners.

In many correlational studies, one of the variables can be regarded as the predictor variable with the other one as the outcome variable. The predictor variable can be seen as occurring before the outcome variable in some sense. It is called the predictor variable, because it forms the basis for predicting the value of the outcome variable. For example, there is a positive correlation between the Type A Behaviour Pattern (hostility, impatience, tension) and coronary heart disease (Miller et al., 1991). Here the Type A Behaviour Pattern is the predictor variable and coronary heart disease is the outcome variable. This approach may suggest the existence of a causal relationship. However, it is very important to remember that “correlations cannot prove causes”. There could be a third factor (e.g. genetic vulnerability) that leads to the Type A Behaviour Pattern and to susceptibility to heart disease.

**Careful measurement**

Another feature of good correlational studies is that the variables are carefully measured. Let us consider an example. Martin et al. (1989) argued that Type A individuals are much more highly motivated than Type B individuals, who are relaxed, patient, and calm. This suggests that Type A individuals might have better job performance than Type Bs; thus, there should be a positive correlation between Type A and job performance. How can job performance be measured? In the case of managers whose jobs involve forward planning, motivating their staff, monitoring the performance of their staff, and so on, it may be very hard to assess their work performance in a single measure. It would be preferable to study a group such as insurance salespeople. Their main work goal is to sell as much insurance as possible, and so the amount of insurance sold over a given period of time (e.g. three months) would provide a reasonable measure of job performance.

**Wide range**

A final feature of good correlational studies is that the scores on both variables vary considerably from individual to individual. For example, IQ is supposed to reflect general intellectual ability, and so one might predict that there would be a positive correlation
between IQ and job performance. This has been found numerous times (Eysenck, 1994). Suppose, however, that we correlated IQ and job performance among chartered accountants. The great majority of chartered accountants have high IQs, and so we have what is known as restriction of range. This restriction of range would reduce the strength of the association between IQ and job performance, and so it should be avoided.

**PROBLEMS WITH EXPERIMENTAL RESEARCH**

In most experimental research (and some non-experimental research), the experimenter and the participants interact with each other. This can produce various kinds of problems. The ways in which experimenters behave and talk may influence the behaviour of the participants in ways that have nothing to do with the independent variable or variables being manipulated. In addition, the participants may form mistaken ideas of what the experiment is about, and these mistaken ideas may affect their behaviour. Some of the main problems stemming from the relationship between the researcher and the participants are discussed in this section.

**Experimenter effects**

The ideal experimenter is someone who behaves in exactly the same mildly positive way with every participant, and who does not allow his or her expectations and experimental hypotheses to influence the conduct of a study. In reality, the experimenter’s expectations, personal characteristics, and so on often have an effect on the participants’ behaviour; these are known as experimenter effects.

**Experimenter expectancy**

One of the most important experimenter effects is experimenter expectancy, in which the experimenter’s expectations have a systematic effect on the performance of the participants. Perhaps the first systematic demonstration of experimenter expectancy involved a horse known as Clever Hans. The horse was apparently able to count, tapping its hoof the right number of times when asked a simple mathematical question (e.g. $8 + 6$). Pfungst (1911) studied Clever Hans. He found that Clever Hans could not produce the correct answer when the horse was blindfolded. What happened normally was that the experimenter made slight movements when the horse had tapped out the correct number, and Clever Hans was simply using these movements as the cue to stop tapping.

**Other effects**

Barber (1976) argued that there are numerous ways in which the experimenter can influence the findings obtained. In addition to experimenter expectancy, he identified several other kinds of experimenter effects (summarised in Coolican, 1994). These effects are listed here. In this list, a distinction is drawn between the investigator (the person directing the research) and the experimenter (the person actually carrying out the experiment). For example, an academic psychologist will often be the investigator, whereas an undergraduate or postgraduate student is the experimenter. So far as the studies carried out by undergraduate studies are concerned, the investigator and the experimenter will typically be the same person.

1. **Investigator paradigm effect**: the entire approach adopted by the investigator can make it harder or easier to obtain certain findings.
2. Investigator experimental design effect: for example, if an investigator wanted to show that a learning programme for disadvantaged children was not effective, he or she could arrange for the programme to last for a very short period of time to reduce the chances that there would actually differ in activity level.

What do you think Rosenthal found? Somewhat surprisingly, the experimenters reported twice as many head turns and three times as many body contractions in the worms that were allegedly “highly active” as in the “inactive” ones! Rosenthal argued that this was an experimenter expectancy effect, but it is more likely that it was due to the experimenters failing to follow the proper procedures and/or misrecording of the data. As Coolican (1994) pointed out, there was no evidence of expectancy effect in at least 40 experiments specifically designed to find it. There is evidence that the behaviour of human participants, especially those high in need for approval, can be influenced by the experimenter’s behaviour. However, it seems less likely that flatworms would respond to a smile or a frown from the experimenter!

Discussion points
1. Are you surprised that it has proved hard to replicate Rosenthal’s findings on flatworms?
2. In what circumstances would you expect to find experimenter effects (see later)?

Reducing experimenter effects

What steps can be taken to minimise experimenter effects? One approach is to use a double blind procedure, in which neither the experimenter working with the participants nor the participants know the experimental hypothesis (or hypotheses) being tested. The double blind procedure reduces the possibility of experimenter bias, but it is often too expensive and impractical to use. However, the incidence of experimenter effects is probably less than it used to be, for the simple reason that more and more experiments involve participants

KEY STUDY EVALUATION — Rosenthal

Psychological experiments like Rosenthal’s are carried out by humans on humans. As such they are unique social situations in which social interactions play an important part. Inevitably, problems can arise in the form of experimenter effects. According to Rosenthal, the participants could be influenced to expect certain results to occur within the experiment. However, it is possible that the experimenter could have given clues as to how the participants were expected to behave, either verbal or non-verbal in nature. Rosenthal suggested that the perceived competence and authority of the research could also produce experimenter effects. This would be influenced by the participant’s own personal characteristics such as a need for approval.

How could the results obtained by Rosenthal be due to the participants deviating from the standardised procedure, rather than to the experimenter expectancy effect?

KEY TERM

Double blind: a procedure in which neither the experimenter nor the participants know the precise aims of the study; where possible, they do not know the condition in which each participant has been placed.
interacting with computers rather than with human experimenters. In addition, data are increasingly stored directly in computers, making it harder to misrecord the information obtained from participants.

**Demand characteristics**

A common criticism of laboratory research is that the situation is so artificial that participants behave very differently from the way they do normally. Guy Claxton (1980) discussed an amusing example of this. He considered a laboratory task, in which participants have to decide as rapidly as possible whether sentences such as “Can canaries fly?” are true or false. Under laboratory conditions, people perform this task uncomplainingly. However, as Claxton pointed out, “If someone asks me ‘Can canaries fly?’ in the pub, I will suspect either that he is an idiot or that he is about to tell me a joke.”

Why do people behave in unusual ways under laboratory conditions? The American psychologist Orne (1962) emphasised the importance of what he termed demand characteristics, which are “the totality of cues which convey an experimental hypothesis to the subjects [participants].” Demand characteristics include “the rumours or campus scuttlebutt about the research, the information conveyed during the original situation, the person of the experimenter, and the setting of the laboratory, as well as all explicit and implicit communications during the experiment proper.” (In case you are wondering, the word “scuttlebut” means gossip.) Orne’s basic idea is that most participants do their best to comply with what they perceive to be the demands of the experimental situation, but their perception will often be inaccurate.

As Orne showed, the demand characteristics in an experiment are so powerful that the participants can often be persuaded to do some very strange things. He discussed one study in which the participants spent several hours adding numbers on random number sheets, then tearing up each completed sheet into at least 32 pieces. Many of the participants treated the situation as a test of endurance, and this motivated them to keep going.

There is another problem with demand characteristics, which applies to participants who have previously taken part in an experiment in which they were deceived about the experimental purpose. As a result of being deceived, some participants tend thereafter to respond in the opposite direction to the one suggested by an experiment’s demand characteristics. Why should this be so? Silverman, Shulman, and Wiesenthal (1970) explained this effect in the following way:

Deceived subjects [participants] may have become so alerted to possible further deceptions that they tend to respond counter to any cues regarding the experimenter’s hypothesis. An element of gamesmanship may enter the experimental situation in that subjects [participants] become wary of “tricks” underlying the obvious, and do not want to be caught in them.

**Reducing demand characteristics**

Information about the demand characteristics in any given experimental setting can be obtained by asking the participants afterwards to describe in detail what they felt the experiment was about. Armed with this information, the experimenter can take steps to make sure that the results of future experiments are not adversely affected by demand characteristics.

Some (but not all) of the problems of demand characteristics can be reduced by the double blind procedure described earlier. Another possibility in some studies is the single blind procedure, in which the participants are not informed of the condition in which they have been placed. However, this raises ethical issues, because full informed consent cannot be obtained in such circumstances.

**Evaluation apprehension**

Rosenberg (1965) pointed out that an important aspect of most participants’ behaviour in the experimental or laboratory situation is what he called evaluation apprehension.
He defined this as “an active anxiety-toned concern that he [the participant] win a positive evaluation from the experimenter or at least that he provide no grounds for a negative one.” It could be argued that the main reason why participants comply with the demand characteristics of experimental situations is because of their evaluation apprehension. However, evidence that the need for favourable personal evaluation can be more important than the need to comply with demand characteristics was reported by Sigall, Aronson, and Van Hoose (1970).

Sigall et al. carried out an experiment on copying telephone numbers. The experimenter told the participants doing the test for the second time that he expected them to perform it at a rate that was actually slower than their previous performance. Adherence to the demand characteristics would have led to slow performance, whereas evaluation apprehension and the need to be capable would have produced fast times. In fact, the participants performed faster than they had done before, indicating the greater importance of evaluation apprehension than of demand characteristics.

GENERAL ISSUES IN INVESTIGATIONS

So far in this chapter, we have considered several specific issues that are important to ensure that the design of a study is appropriate. In this section, we will address some important general criteria that can (and should) be used to evaluate how successfully a study has been designed and carried out. The criteria to be discussed are participant reactivity; validity; generalisability; and reliability.

Participant reactivity

A weakness that is found in many studies is what is known as participant reactivity. This refers to a situation in which an independent variable has an effect on behaviour simply because the participants know that they are being observed or studied. Any measure of the participants’ behaviour which could suffer from this effect is called a reactive measure, and reactivity is the term used to refer to the changes in behaviour produced in this way.

Validity and generalisability

One of the key requirements of a study or experiment is that any findings obtained are valid, in the sense that they are genuine and provide us with useful information about the phenomenon being studied. Campbell and Stanley (1966) drew a distinction between internal validity and external validity, which is of most relevance to experiments and quasi-experiments. Internal validity refers to the issue of whether the effects observed are genuine and are caused by the independent variable. In contrast, external validity refers to the extent to which the findings of a study can be generalised to situations and samples other than those used in the study. This distinction between two kinds of validity is an important one: many experiments possess internal validity while lacking external validity (see the Research methods: Psychological enquiry chapter).

Internal validity

We will shortly consider some of the reasons why an experiment may lack external validity, but what are some of the main threats to the internal validity of an experiment? Coolican (1994) pointed out that there are many such threats, most of which were discussed earlier in the chapter. For example, the existence of any confounding factors threatens internal validity, as does the use of unreliable or inconsistent measures. Problems with internal validity can also arise if an experiment is designed without careful attention being paid to issues such as standardisation, counterbalancing, and randomisation. Other threats to internal validity include experimenter effects, demand characteristics, participant reactivity, and the use of inappropriate statistical tests. In a nutshell, virtually all of the principles of experimental design are intended to enhance internal validity, and failures to apply these principles threaten internal validity. If internal validity is high, then...
there are good prospects for being able to replicate the findings. If it is low, then replication is likely to be difficult or impossible.

**External validity and generalisability**

What about external validity? There are close links between external validity and **generalisability**, because both are concerned with the issue of whether the findings of an experiment or study are applicable to other situations. More specifically, Coolican (1994) argued that there are four main aspects to external validity or generalisability, which we consider in turn:

- **Populations**: do the findings obtained from a given sample of individuals generalise to a larger population from which the sample was selected?
- **Locations**: do the findings of the study generalise to other settings or situations? If the findings generalise to various real-life settings, then the study is said to possess ecological...
validity. Silverman (1977, p. 108) was sceptical about the ecological validity of laboratory experiments: “the conclusions we draw from our laboratory studies pertain to the behaviour of organisms in conditions of their own confinement and control and are probably generalisable only to similar situations (institutions, perhaps, such as schools or prisons).”

- **Measures or constructs**: do the findings of the experiment or study generalise to other measures of the variables used? For example, suppose we find that people who are high on the personality dimension of trait anxiety as assessed by Spielberger’s State–Trait Anxiety Inventory have worse long-term memory measured by recall than those low in trait anxiety. Would we obtain the same findings if trait anxiety were assessed by a different questionnaire or if we used a recognition test of long-term memory?

- **Times**: do the findings generalise to the past and to the future? For example, it could be argued that the sweeping changes in many cultures in recent decades have affected conformity behaviour as studied by Asch, and obedience to authority as studied by Milgram (see PIP, Chapter 20).

What can we do to maximise the external validity of an experiment? Unfortunately, there is no easy answer to that question. What usually happens is that the external validity of an experiment only becomes clear when other researchers try to generalise the findings to other samples or populations, locations, measures, and times. It might be thought that the findings of field experiments are more likely than those of laboratory experiments to generalise to other real-life locations or settings, but that is not necessarily so.

**Meta-analyses.** One way of trying to determine whether certain findings generalise is to carry out what is known as a *meta-analysis*. What is done in a meta-analysis is to combine all of the findings from many studies designed to test a given hypothesis into a single analysis. If the meta-analysis indicates that some finding has been obtained consistently, this suggests that it generalises across populations, locations, measures, and times. For example, Smith et al. (1980) discussed a meta-analysis on over 400 studies concerned with the effectiveness of psychotherapy. They concluded that psychotherapy was reasonably effective, because patients receiving psychotherapy improved more than did 75% of the patients not receiving any therapy.

The greatest limitation of meta-analyses is that differences in the quality of individual studies are often ignored. This can lead to the situation in which a finding is accepted as genuine when it has been obtained in several poorly designed studies but not in a smaller number of well-designed studies. Another problem is that it is often hard to know which studies to include and which to exclude. For example, Smith et al. considered all forms of non-behavioural therapy together. However, some forms of non-behavioural therapy were more effective than others (Barlow & Durand, 1995), so it was perhaps undesirable to put them together into a single meta-analysis.

**Key Terms**

**Meta-analysis**: an analysis in which all of the findings from many studies relating to a given hypothesis are combined for statistical testing.
Reliability

One of the main goals of experimental research is to design and carry out studies in such a way that replication or repetition of its findings is possible. In order to achieve that goal, it is important that the measures we use should possess good reliability or consistency. As Coolican (1994, p. 50) pointed out:

*Any measure we use in life should be reliable, otherwise it’s useless. You wouldn’t want your car speedometer or a thermometer to give you different readings for the same values on different occasions. This applies to psychological measures as much as any other.*

Problems relating to reliability are likely to arise when the experimenter is trying to code the complex behaviour of participants using a manageable number of categories. For example, it is common in studies of naturalistic observation to record certain events (e.g. performing an aggressive act). However, it may be hard to define those events with enough precision to produce reliable results. One way of assessing this is by asking two (or more) judges to provide ratings in the observational situation. The ratings can then be compared to provide a measure of inter-judge reliability.

**PERSONAL REFLECTIONS**

- As you were reading this chapter, you may have thought that the various recommendations for designing experimental and non-experimental studies seemed fairly obvious and easy to follow in practice. In fact, it is very hard to design a study that has no flaws. Three of the greatest problems with experimental designs occur with respect to operationalisation, experimenter effects, and external validity. Thus, it is usually difficult to find ways of operationalising key variables; to avoid the experimenter influencing the participants’ behaviour in unwanted ways; and to ensure that the findings will apply to other situations and participants.

**SUMMARY**

**Aims and hypotheses**

The first stage in designing a study is to decide on its aims and hypotheses. There will generally be an experimental or alternative hypothesis and a null hypothesis. The experimental hypothesis may be directional or one-tailed, or it may be non-directional and two-tailed.

**Selecting participants**

The participants selected for a study represent a sample from some population. They should form a representative sample; in other words, sampling bias should be avoided. The best approach is random sampling, but other reasonable methods are systematic sampling, stratified sampling, and quota sampling. Opportunity sampling is the easiest but least satisfactory method. The sample size depends on the likely size of the effect being studied.

**Good practice in experimentation**

It is important to use standardised procedures. It is also important to avoid confounding variables and other forms of constant error, and to keep random error to a minimum. Operationalisation is useful, but operational definitions typically cover only part of the meaning of the independent or dependent variable in question.

**Experimental designs**

There are three main types of experimental design: independent design; matched participants design; and repeated measures design. With an independent design, randomisation...
is generally used to allocate the participants to groups. Counterbalancing is often used with the repeated measures design in order to balance out any order effects and prevent them from disrupting the findings.

We can distinguish between participant observation and non-participant observation research. Participant research involves the three stages of entering the field; being in the field; and leaving the field. Non-participant observation research involves devising precise, comprehensive, and usable behavioural categories. The sampling of behaviour can be continuous; based on time intervals; based on time points; or random. Survey studies can use various designs: one-shot; before–after; two-groups controlled comparison; two-groups before–after. When questionnaires or attitude scales are constructed, the items need to be short, unambiguous, and unbiased, and the tests need to be reliable and valid. Correlational studies should be based on an underlying theory, the variables should be carefully measured, and the scores on both variables should vary considerably from individual to individual.

Most research involves interactions between the experimenter and the participants. This can introduce various systematic biases, which can be divided into experimenter effects and demand characteristics. Experimenter effects include experimenter expectancy, experimenter misrecording, and experimenter fudging. Demand characteristics involve the participants responding on the basis of their beliefs about the experimental hypothesis or hypotheses. In addition, the behaviour of participants is sometimes influenced by evaluation apprehension.

In some studies, the independent variable has an effect on behaviour simply because the participants know they are being observed. This is known as participant reactivity or the Hawthorne effect. It is a serious problem, because it can lead us to misinterpret our findings. It is important for a study to have internal validity, meaning that the findings are genuine and caused by the independent variable. External validity, which refers to the extent to which the findings of a study can be generalised, is also important. Issues of generalisability apply to populations, locations, measures, and times. Information about the generalisability of any particular findings can be obtained by means of a meta-analysis. The measures used in a study should possess good reliability or consistency. If they do not, then they are inadequate measures of the variables in question, and it will be hard to replicate or repeat any findings obtained.

**Further Reading**


**Revision Questions**

Sample questions on research methods are given at the end of the Research methods: Data analysis chapter.

**References**


