Chapter 16
Psychological research and scientific method

SPECIFICATION

This section builds on the knowledge and skills developed at AS level.

Psychological research and scientific method	
The application of scientific method in psychology	• The major features of science, for example replicability, objectivity. • The scientific process, including theory construction, hypothesis testing, use of empirical methods, generation of laws/principles (e.g. Popper, Kuhn). • Validating new knowledge and the role of peer review.
Designing psychological investigations	• Selection and application of appropriate research methods. • Implications of sampling strategies, e.g. bias and generalising. • Issues of reliability, including types of reliability, assessment of reliability, improving reliability. • Assessing and improving validity (internal and external). • Ethical considerations in design and conduct of psychological research.
Data analysis and reporting on investigations	• Appropriate selection of graphical representations. • Probability and significance, including the interpretation of significance and Type 1/Type 2 errors. • Factors affecting choice of statistical test, including levels of measurement. • The use of inferential analysis, including Spearman's *Rho*, Mann-Whitney, Wilcoxon, chi-square. • Analysis and interpretation of qualitative data. • Conventions of reporting on psychological investigations.

The scientific process has been inherent in all the psychology you have studied. Psychologists, like all scientists, conduct empirical, objective research wherever possible in order to test hypotheses and construct theories. For example, you have undoubtedly come across **social learning theory** during your study of psychology. This theory explains how we learn new behaviours through indirect as well as direct reinforcement. It developed as a result of various research studies such as the well-known Bobo doll experiments (see page 60). These are an example of the scientific method or 'process', described in the box at the bottom of this page.

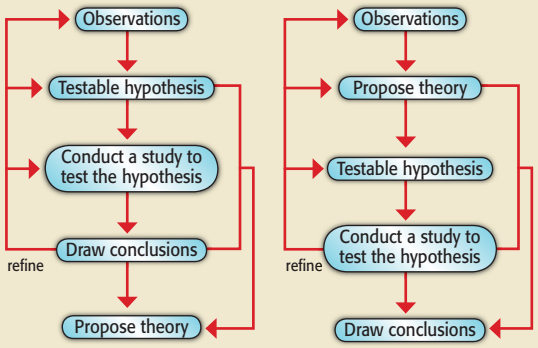
Science is: 'A branch of knowledge conducted on objective principles involving the systematised observation of and experiment with phenomena.' (Oxford Concise Dictionary)

AN EMPIRICAL TEST

On the left is a picture of a burger from a well-known fast-food outlet. This is what you are led to expect you will get. But what about reality? You may think you know something, but unless you test this empirically you cannot know if it is true. On the right is the **empirical** evidence of what the burgers are really like. 'Empirical' refers to information gained through direct observation. Science uses empirical methods to separate unfounded beliefs from real truths.

[Thanks to Professor Sergio della Sala of Edinburgh University for this tasty and memorable example of empiricism.]

THE SCIENTIFIC PROCESS



▲ Induction ▲ Deduction

The scientific process starts with observations of phenomena in the world. In the **inductive model** this leads scientists to develop hypotheses. Hypotheses are tested, possibly leading to new questions and new hypotheses. Eventually such data may be used to construct a theory.

The **deductive model** places theory construction at the beginning, after making observations.

A 'good' theory is one that can be empirically tested. Unless you can test a theory there is no means of knowing if it is right or wrong. A good theory should therefore produce a variety of testable hypotheses, thus allowing falsification.

WHAT IS SCIENCE?

The major features of science

Science is a means of finding out about our world i.e. gaining knowledge. However, most importantly, it aims to uncover facts that can be relied on. Such knowledge enables us to control our world, for example build bridges and treat schizophrenia. If the knowledge is not true, our bridges will collapse and our treatments won't work.

In order to uncover 'truths' about the world, scientists use the scientific method. The key features of this method are:

- **Empiricism** – Information is gained through direct observation or experiment rather than by reasoned argument or unfounded beliefs.
- **Objectivity** – Scientists strive to be objective in their observations and measurements i.e. their expectations should not affect what they record.
- **Replicability** – One way to demonstrate the validity of any observation or experiment is to repeat it. If the outcome is the same, this affirms the truth of the original results, especially if the observations have been made by a different person. In order to achieve such replication, it is important for scientists to record their methods carefully so that the same procedures can be followed in the future.
- **Control** – Scientists seek to demonstrate causal relationships to enable them to predict and control our world. The experimental method is the only way to do this – where we vary one factor (the **independent variable**) and observe its effect on a **dependent variable**. In order for this to be a 'fair test' all other conditions must be kept the same, i.e. controlled.
- **Theory construction** – One aim of science is to record facts, but an additional aim is to use these facts to construct **theories** to help us understand and predict the natural phenomena around us. A theory is a collection of general principles that explain observations and facts.

The scientific process

The diagram on the left represents one model of the scientific process (or scientific method). Theory construction may occur at the beginning or end of the process:

1 ***Induction*** involves reasoning from the particular to the general. For example, a scientist may observe instances of a natural phenomenon and come up with a general law or theory. Before the twentieth century, science largely used the principles of induction – making discoveries about the world through accurate observations, and formulating theories based on the regularities observed. Newton's Laws are an example of this. He observed the behaviour of physical objects and produced laws that made sense of what he observed.

2 ***Deduction*** involves reasoning from the general to the particular, starting with a theory and looking for instances that confirm this. Darwin's theory of evolution is an example of this. He formulated a theory and set out to test its propositions by observing animals in nature. He specifically sought to collect data to prove his theory. The **hypothetico-deductive model** was proposed by Karl Popper (1935), suggesting that theories/laws about the world should come first and these should be used to generate expectations/hypotheses which can be **falsified**. Falsification is the only way to be certain – as Popper pointed out: 'No amount of observations of white swans can allow the inference that all swans are white, but the observation of a single black swan is sufficient to refute that conclusion'.

THE APPLICATION OF THE SCIENTIFIC METHOD IN PSYCHOLOGY

Can psychology claim to be a science?

Scientific research is desirable – In the nineteenth century, early psychologists sought to create a *science* of psychology because this would enable them to produce verifiable knowledge as distinct from common sense or 'armchair psychology'. We could claim that men are more aggressive than women, or that ECT cures depression, but people would be likely to demand proof of such claims.

Psychology is a science insofar as it shares the goals of all sciences and uses the scientific method. Most psychologists generate models which can be falsified, and conduct well-controlled experiments to test these models. However, there is the question of whether simply using the scientific method turns psychology into a science. Miller (1983) suggests that psychologists who attempt to be scientists are doing no more than 'dressing up'. They may take on the tools of sciences such as quantified measurements and statistical analysis but the essence of science has eluded them. Perhaps at best it is a **pseudoscience** – but it is a dangerous one, because psychologists can then claim that their discoveries are fact.

Kuhn's views – Thomas Kuhn (1962) claimed that psychology could not be a science because, unlike other sciences, there is no single **paradigm** (i.e. a shared set of assumptions). A science such as biology or physics has a unified set of assumptions, whereas psychology has a *number* of paradigms or approaches – cognitive, physiological, behaviourist, evolutionary, psychoanalytic and so on. Therefore Kuhn suggested psychology was a 'pre-science'.

Lack of objectivity and control – Some psychologists claim that human behaviour can be measured as objectively as the measurement of physical objects (objectivity is a key goal of science). But is this true? In psychology the object of study reacts to the researcher and this leads to problems such as **experimenter bias** and **demand characteristics**, which compromise validity. However, similar problems apply to the hard sciences. Heisenberg (1927) argued that it is not even possible to measure a subatomic particle without altering its 'behaviour' in doing the measurement. This *uncertainty principle* is a kind of experimenter effect: the presence of an experimenter changes the behaviour of what is observed, even in physics.

Are the goals of science appropriate for psychology?

Some psychologists do not see the study of behaviour as a scientific pursuit. For example the psychiatrist R.D. Laing (1965), in discussing the causes of schizophrenia, claimed that it was inappropriate to view a person experiencing distress as a complex physical-chemical system that had gone wrong. Laing claimed that treatment could only succeed if each patient was treated as an individual case (the **idiographic** approach). Science takes the **nomothetic** approach, looking to make generalisations about people and find similarities. Perhaps the way to decide whether science is appropriate for psychology is to look at the results of psychological research – psychological approaches to treating mental illness have had at best modest success, which suggests that the goals of science are not always appropriate.

Qualitative research – Some psychologists advocate more subjective, qualitative methods of conducting research (see page 298). However these methods are still 'scientific' insofar as they aim to be valid. For example, data can be collected from interviews, discourse analysis, observations, etc. and then **triangulated** – the findings from these different methods are compared with each other as a means of verifying them and making them objective.

REDUCTIONIST AND DETERMINIST

The scientific approach is both reductionist and determinist. It is **reductionist** because complex phenomena are reduced to simple variables in order to study the causal relationships between them. It is also reductionist in the development of theories – the *canon of parsimony* or *Occam's razor* (a principle attributed to the mediaeval philosopher William of Occam) states that: 'Of two competing theories or explanations, all other things being equal, the simpler one is to be preferred'.

Science is also **determinist** in its search for causal relationships, i.e. seeking to discover if X determines Y. If we don't take a determinist view of behaviour, this rules out scientific research as a means of understanding behaviour.

Reductionism and determinism are mixed blessings. If we reduce complex behaviour to simple variables this may tell us little about 'real' behaviour, and yet without this reductionism it is difficult to pick out any patterns or reach conclusions. Determinism may also oversimplify the relationship between causes and effects but provides insights into important factors, such as the influences of nature and nurture.

PROVE IT

Some years ago this photograph was doing the rounds on the internet. Apparently a crew member from an oil rig took a nap during the middle of his shift and didn't reappear. When his fellow workers went searching for him they came upon a python who had clearly eaten a large meal. They captured and killed the python and upon opening it up, they found their friend.

Why are we all so gullible? People *want* to believe such stories (this, incidentally, is a rigged photo). In this case there are no serious consequences of 'believing' but there are many situations where people are sweet-talked into spending large sums on a miracle cure or fooled by newspapers reporting a 'scientific proof'. The only protection against charlatans and pseudoscience is an understanding of the goals of science and having a permanently sceptical mind.

CAN YOU...? No.16.1

...1 Select **two** features of science and explain why each is important as a means of gaining 'true' knowledge.

...2 Outline the scientific process (either induction or deduction).

...3 Briefly explain the views of Popper and Kuhn.

...4 Briefly discuss, in about 150 words, the application of the scientific method to psychology.

...5 Select **one** theory you have studied during your Psychology course and consider whether it is an example of induction or deduction.

...6 Using the model of the scientific process, explain the construction of the theory that you selected in question 5, including initial observations and the process of hypothesis testing.

VALIDATING NEW KNOWLEDGE

Psychology, in common with all scientific subjects, develops its knowledge base through conducting research and sharing the findings of such research with other scientists. **Peer review** is an essential part of this process and scientific quality is judged by it. It is in the interest of all scientists that their work is held up for scrutiny and that any work that is flawed or downright fraudulent (as in 'The Cyril Burt Affair' – see facing page), is detected and its results ignored.

SCHOLARLY JOURNALS

There are thousands of scholarly journals, publishing over a million research papers each year. They differ from 'popular' magazines because they contain in-depth reports of research. The articles are written by academics and are peer-reviewed. Several hundred such journals specifically relate to psychology, such as *The Psychologist, Archives of Sexual Behaviour*, *Journal of Early Adolescence* and *British Journal of Psychology.*

These journals may be published weekly, monthly or less frequently. Academic textbooks are based on articles published in the journals and link research claims to scholarly reports – as you can see by looking at the references in the back of this book.

THE ROLE OF PEER REVIEW

Peer review (also called 'refereeing') is the assessment of scientific work by others who are experts in the same field (i.e. 'peers'). The intention of peer reviewing is to ensure that any research conducted and published is of high quality.

Peer reviewers are generally unpaid. Usually there are a number of reviewers for each application/article/assessment. Their task is to report on the quality of the research and then their views are considered by a peer review panel.

The Parliamentary Office of Science and Technology (2002) suggests that peer review serves three main purposes:

1. **Allocation of research funding** – Research is paid for by various government and charitable bodies. The government-run Medical Research Council (MRC), for example, is one of the leading UK sources of funding for research. In 2008–9 it had £605 million to spend (Hansard, 2007) and obviously a duty to spend this responsibly. Therefore public bodies such as the MRC require reviews to enable them to decide which research is likely to be worthwhile.
2. **Publication of research in scientific journals and books** – Scientific or scholarly journals provide scientists with the opportunity to share the results of their research. The peer review process has only been used in such journals since the middle of the twentieth century as a means of preventing incorrect or faulty data entering the public domain. Prior to peer review, research was simply published and it was assumed that the burden of proof lay with opponents of any new ideas.
3. **Assessing the research rating of university departments** – All university science departments are expected to conduct research and this is assessed in terms of quality (Research Assessment Exercise, RAE). Future funding for the department depends on receiving good ratings from the RAE peer review.

Peer review and the internet

The sheer volume and pace of information available on the internet means that new solutions are needed in order to maintain the quality of information. Scientific information is available in numerous online blogs, online journals and, of course, *Wikipedia* (an online encyclopedia). To a large extent such sources of information are policed by the 'wisdom of crowds' approach – readers decide whether it is valid or not, and post comments and/or edit entries accordingly. Several online journals (such as *ArXiv* and *Philica*) ask readers to rate articles. On Philica, papers are ranked on the basis of peer reviews. On the internet, however, 'peer' is coming to mean 'everyone' – perhaps a more egalitarian system.

CONVENTIONS FOR REPORTING PSYCHOLOGICAL INVESTIGATIONS

Scientific journals contain research reports which tend to be organised into the following sections:

Abstract – A summary of the study covering the aims/hypothesis, method/procedures, results and conclusions.

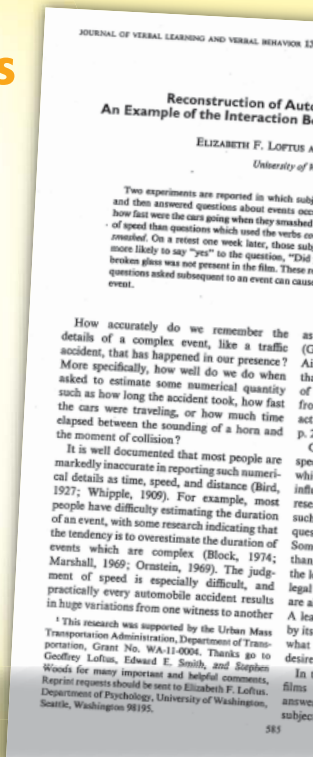
Introduction/Aim – What the researchers intend to investigate. This often includes a review of previous research (theories and studies), explaining why the researchers intend to conduct this particular study. The researchers may state their research predictions and/or a hypothesis or hypotheses.

Method – A detailed description of what the researchers did, providing enough information for replication of the study. Included in this section is information about the participants (the sample), the testing environment, the procedures used to collect data, and any instructions given to participants before (the brief) and afterwards (the debrief).

Results – This section contains what the researchers found, often called statistical data, which includes descriptive statistics (tables, averages and graphs) and inferential statistics (the use of statistical tests to determine how significant the results are).

Discussion – The researchers offer explanations of the behaviours they observed and might also consider the implications of the results and make suggestions for future research.

References – The full details of any journal articles or books that are mentioned.



Reconstruction of Aut
An Example of the Interaction B

ELIZABETH F. LOFTUS

► This is the first page of the journal article by Loftus and Palmer (1974) on the effect of leading questions. Have a look at some journal articles yourself by ordering them from the Inter-Library loan system at your local library.

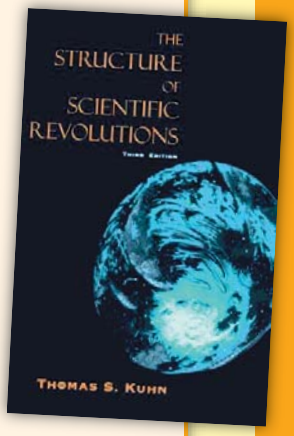
COMMENTARY

It is clear why peer review is essential – without it we don't know what is mere opinion and speculation, and what is real fact. We need to have a means of establishing the validity of scientific research.

While the purpose of peer review is beyond question, certain features of the process can be criticised. For example Richard Smith, previous editor of the *British Medical Journal (BMJ)* commented: 'Peer review is slow, expensive, profligate of academic time, highly subjective, prone to bias, easily abused, poor at detecting gross defects and almost useless at detecting fraud' (Smith, 1999). Let us pick up a few of these criticisms.

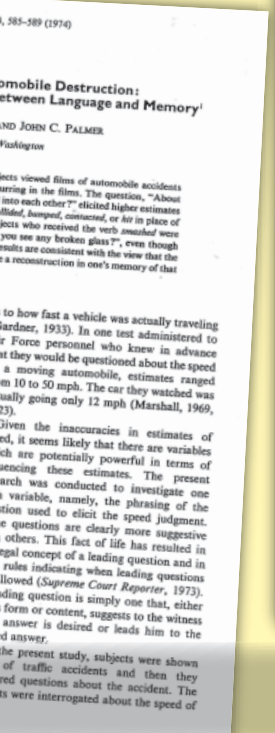
- **Unachievable ideal** – It isn't always possible to find an appropriate expert to review a research proposal or report. This means that poor research may be passed because the reviewer didn't really understand it.
- **Anonymity** is usually practised so that reviewers may be honest and objective. However it may have the opposite effect if reviewers use the veil of anonymity to settle old scores or bury rival research. Research is conducted in a social world where people compete for research grants and jobs, and make friends and enemies. Social relationships inevitably affect objectivity. Some journals now favour open reviewing (both author and reviewer know each other's identity).
- **Publication bias** – Peer review tends to favour the publication of positive results, possibly because editors want research that has important implications in order to increase the standing of their journal.
- **Preserving the status quo** – Peer review results in a preference for research that goes with existing theory rather than dissenting or unconventional work. The former editor of the medical journal, *The Lancet*, Richard Horton (2000), made the following comment: 'The mistake, of course, is to have thought that peer review was any more than a crude means of discovering the acceptability – not the validity – of a new finding'. This is in accord with Kuhn's views about scientific revolutions (see right) – science is generally resistant to large shifts in opinion. Change takes a long time and requires a 'revolution' in the way people think. Peer review may be one of the elements that slows down change.

HOW DOES KNOWLEDGE CHANGE?



Thomas Kuhn published an enormously influential book called *The structure of scientific revolutions* (Kuhn, 1962). He proposed that scientific knowledge about the world develops through revolutions, rather than the process suggested by Popper's theory of falsification whereby theories are fine-tuned by a successive series of experiments. Kuhn proposed that there are two main phases in science. One is called 'normal science', where one theory remains dominant, despite occasional challenges, by disconfirming studies. Gradually the disconfirming evidence accumulates until the theory can no longer be maintained, and then it is overthrown. This is the second phase – a revolutionary shift. Kuhn didn't use the term 'theory' – he spoke of a **paradigm**, which he defined as 'a shared set of assumptions about the subject matter of a discipline and the methods appropriate to its study'. A science (like physics or biology) has a unified set of assumptions and methods.

A classic example of a paradigm shift was the revolution in our understanding of the universe due to the work of the Polish astronomer Copernicus in the sixteenth century. Copernicus overthrew the belief held for almost 2000 years that the earth was the centre of the universe. Such changes are not logical, as Popper's view of science might suggest. According to Kuhn, scientific progress is more like a religious conversion and is related to social factors. Kuhn's view itself is potentially an example of a paradigm shift in the sciences – a change of view from science as logic to science as a **social construction** – though at present this view is still hotly debated (Jones and Elcock, 2001).


585–589 (1974)

omobile Destruction:
etween Language and Memory¹

ND JOHN C. PALMER
Washington

jects viewed films of automobile accidents
urring in the films. The question, "About
into each other?" elicited higher estimates
llided, bumped, contacted, or *hit* in place of
jects who received the verb *smashed* were
you see any broken glass?", even though
sults are consistent with the view that the
a reconstruction in one's memory of that

to how fast a vehicle was actually traveling
ardner, 1933). In one test administered to
r Force personnel who knew in advance
t they would be questioned about the speed
a moving automobile, estimates ranged
en 10 to 50 mph. The car they watched was
ually going only 12 mph (Marshall, 1969,
23).
Given the inaccuracies in estimates of
ed, it seems likely that there are variables
ch are potentially powerful in terms of
encing these estimates. The present
arch was conducted to investigate one
variable, namely, the phrasing of the
tion used to elicit the speed judgment.
e questions are clearly more suggestive
others. This fact of life has resulted in
egal concept of a leading question and in
rules indicating when leading questions
llowed (*Supreme Court Reporter*, 1973).
ding question is simply one that, either
form or content, suggests to the witness
answer is desired or leads him to the
d answer.
he present study, subjects were shown
of traffic accidents and then they
red questions about the accident. The
s were interrogated about the speed of

THE CYRIL BURT AFFAIR – AN EXAMPLE OF SCIENTIFIC FRAUD

In the early 1950s, the eminent psychologist Sir Cyril Burt published results from studies of identical twins that was used as evidence to show that intelligence is inherited. Burt (1955) started with 21 pairs of twins, later increasing this to 42 pairs of twins reared apart. In a subsequent study Burt (1966) increased his sample to 53 pairs of identical twins raised apart, reporting an identical correlation of .771 to the earlier twin study. The suspicious consistency of these correlation coefficients led Kamin (1974) to accuse Burt of inventing data. When a reporter, Oliver Gillie (1976), tried but failed to find two of Burt's research assistants this appeared to confirm the underlying fraud and Burt was publicly discredited. These accusations have been challenged (e.g. Joynson, 1989) but the most recent view is that Burt was astonishingly dishonest in his research (Mackintosh, 1995).

The Burt affair is particularly worrying because we trust psychologists to be honest, and also because his research was used to shape social policy. Burt helped to establish the Eleven-Plus examination used in the UK to identify which children were brighter and should go to grammar school rather than secondary moderns. He argued that since IQ was largely genetic, it was appropriate to test and segregate children into schools appropriate to their abilities. The discovery of Burt's fraud may have subsequently contributed to the move away from grammar schools.

CAN YOU…? No.16.2

…1 Explain what is meant by 'peer review'.

…2 Explain why peer review is essential to the process of producing valid scientific data.

…3 Outline **two** criticisms of peer review.

…4 Produce a written report of a research study, following the conventions for reporting research outlined on the facing page. Select a study you are familiar with from your AS or A2 studies. In some places you may need to invent information (such as describing what procedures might have been used or what graphs might have been displayed). You can look at the internet for more details of the study to help you prepare the report.

RESEARCH METHODS AND CONCEPTS

This chapter began with a consideration of science and the scientific process. The ideal form of scientific investigation is the laboratory experiment because it enables us to maximise **control** and identify causal relationships. However experiments, as you know, are by no means the only form of research method used by scientists and psychologists. On this spread we will review the research methods and techniques that you studied at AS level. In addition, we will also review a number of other concepts you studied as part of the AS course. At A2 level you are expected to build on the knowledge and skills developed at AS.

EXAM TIP *In the A2 Unit 4 examination there will be a set of compulsory questions on 'Psychological research and scientific method'. These questions will be worth a total of 35 marks and will probably start with a brief description of a psychological study followed by a number of questions, similar to those on the facing page. There is also an example of such questions at the end of this chapter.*

Experiments

All experiments involve an **IV** (**independent variable**) and **DV** (**dependent variable**). The IV is varied in order to see how this affects the DV, thus demonstrating a causal relationship. As far as possible, all other variables are controlled, so any changes in the DV are due to the IV rather than **extraneous variables**.

Laboratory experiment – An experiment conducted in a controlled environment which therefore tends to be high in terms of **internal validity** because many extraneous variables can be controlled. However, some (such as investigator/experimenter effects and demand characteristics) may reduce internal validity. Control also increases replicability, which is desirable, but reduces **external validity** because a highly controlled situation may be less like everyday life.

Field experiment – An experiment conducted in a more natural environment. It may be possible to control extraneous variables, though such control is more difficult than in a lab experiment. Experimenter effects are reduced because participants are usually not aware of being in a study. However **demand characteristics** may still be problematic, for example the way an IV is operationalised may convey the experimental hypothesis to participants.

Natural experiment – An experiment which makes use of existing IVs, such as a treatment used for people with schizophrenia. Strictly speaking, an experiment involves the deliberate manipulation of an IV by the experimenter, so causal conclusions cannot be drawn from a natural experiment. In addition, participants are not randomly allocated to conditions in a natural experiment, which may reduce validity, but it is often the only way to study certain behaviours or experiences, such as the effects of privation.

Experimental design – In any experiment there are several levels of the IV. For example in the study of eyewitness testimony by Loftus and Palmer (1974) participants were given a sentence with one of five verbs (hit, smashed, bumped, etc) so there were five levels of the IV. Experimenters have a choice – either each participant is tested on all of the IVs (**repeated measures**) or there are separate groups for each IV (**independent groups**). There is also a third possibility – the participants in each independent group can be matched with participants in the other group(s) on key variables such as age and IQ (**matched pairs** design).

Experimental control is a balancing act

Too little control means it is difficult to draw clear conclusions because of extraneous variables – variables other than the IV which may affect the DV. ▶

◀ Too much control means the behaviour we are studying isn't very much like everyday life (lacks mundane realism).

Self-report methods

Psychologists use **questionnaires** and **interviews** to find out what people think and feel. Interviews are essentially real-time, face-to-face (or over the phone) questionnaires, though there is the option to conduct a fairly **unstructured interview** where the questions are developed by the interviewer as a response to the answers given by the interviewee. **Structured** questionnaires/interviews can be more easily repeated in exactly the same way than unstructured interviews, which is an advantage.

The main problem for **self-report methods** is honesty because, for example, the social desirability bias means that respondents may provide answers to put themselves in a good light.

Questionnaires and interviews may involve **open questions** which permit a respondent to provide their own answer. Such questions can produce unexpected answers providing rich insights but they are more difficult to analyse than **closed questions**.

Observational studies

Perhaps one of the most obvious research methods is simply to watch what people do (**observational techniques**). However it is not that simple to observe behaviour because there is so much information to collect. For this reason, psychologists use **behavioural categories** to record particular instances of behaviour, and also **sampling methods** such as recording behaviour every 30 seconds (**time sampling**) or every time a certain behaviour occurs (**event sampling**).

Even in **naturalistic observations** (as distinct from **controlled observations**) structured techniques are used to study behaviour.

Observational studies provide a rich picture of what people actually do (rather than what they say they do). However observers may be biased (**observer bias**) – their observations can be affected by their expectations.

Experimental control is a balancing act

Correlational analysis

Some studies are concerned with the relationship between two variables, such as IQ and A-level results (which we would expect to be positively correlated) or reaction time and age (which might be negatively correlated). Such studies use a **correlational analysis** which does not demonstrate a cause but is useful in identifying where relationships between co-variables exist. Such studies can be done with large data sets and can be easily replicated. However there may be other, unknown (intervening) variables that can explain why the co-variables being studied are linked. Such studies may lack internal/external validity, for example the method used to measure IQ may lack validity or the sample may lack generalisability.

When answering these questions you may also need to refer to your AS notes on research methods as well as the information on this spread.

...1 For each of the research methods named on this spread, identify an example from the research you are familiar with, and for each example explain **one** strength and **one** weakness of using this research method, in the context of your research example.

...2 Give an example of the following terms in the context of a named study: (a) demand characteristics, (b) pilot study, (c) social desirability bias, (d) extraneous variable, (e) longitudinal study. In each case use the study as a way to provide a clear explanation of the concept.

...3 In the box on the right, a few studies have been briefly described. Try to answer the questions below in relation to each of these studies (or use some other studies you have looked at during your Psychology course).

(a) Identify the research method(s) and/or research technique(s) used in this study and explain your choice.

(b) Write a suitable hypothesis for the study.

(c) Identify whether your hypothesis is directional or non-directional and explain why you chose this kind of hypothesis.

(d) If appropriate, identify the type of experimental design used in the study and explain why the researcher would have chosen this experimental design.

(e) Identify **one** possible extraneous variable that could be a problem in this study and explain how it could be dealt with.

...4 A psychologist intends to study the behaviour of car drivers. Suggest **three** behavioural categories that could be used when recording behaviour.

...5 The government plans a new campaign related to speeding. It decides to find out about attitudes towards speeding in order to make the campaign effective. Suggest **one** closed question and **one** open question that might help the researchers find out more about people's attitudes to speeding.

...6 Design a study to investigate the relationship between success at school and attitudes to school. You should include sufficient details to permit replication, for example, a hypothesis, variables being studied, and detail of design and procedures.

RESEARCH STUDIES

Study A – Johnson and Scott (1976) conducted a study of weapon focus (eyewitness identification may be unreliable because witnesses are distracted by the weapon and take less notice of an assailant's face). In this study participants were asked to sit in a waiting room before being called for the experiment. While waiting, the participants heard a man shouting next door and then someone (a confederate) ran through the room who was either holding a pen and had grease on his hands, or held a bloodied letter opener. Participants were then asked to identify the person who had run through the room and were found to be less accurate if the confederate had been holding the knife.

Study B – Buss (1989) explored what males and females looked for in a marriage partner. The study involved over 10,000 people from 37 different cultures. Participants were asked to rate 18 characteristics (e.g. dependability, chastity, intelligence) in terms of desirability in a mate. The results from men and women were compared and they found, for example, that more women than men desired mates who were 'good financial prospects' whereas more men than women placed importance on physical attractiveness.

Study C – Rahe *et al.* (1970) investigated the relationship between stressful life events and illness by studying a large group of men in the US Navy and seeing if there was an association between the number of stressful life events they experienced over a six-month period and the number of illnesses they experienced.

Study D – Kiecolt-Glaser *et al.* (1984) also investigated stress. In this study the effects of short-term stress on the immune system were demonstrated by comparing blood samples taken from students one month prior to their exams (low stress) and during the exam period itself (high stress).

Case studies

A **case study** is a detailed study of a single individual, institution or event. It uses information from a range of sources, such as from the person concerned and also from their family and friends. Many techniques may be used such as interviews, psychological tests, observations and experiments. Case studies are generally **longitudinal**: in other words they follow the individual or group over an extended period of time. The complex interaction of many factors can be studied, in contrast to experiments where many variables are held constant. However it is difficult to generalise from individual cases as each one has unique characteristics. It is often necessary to use recollection of past events as part of the case study and such evidence may be unreliable.

Other research methods

There are numerous other methods such as **content analysis**, a kind of observational study; **cross-cultural research**, comparing the effects of different cultural practices on behaviour; and **meta-analysis**, which combines the results of many studies on the same topic to reach overall conclusions.

A FEW OTHER CONCEPTS

Aims and hypotheses – Researchers start by identifying what they intend to study (the **aims**) and then make a formal statement of their expectations using a **hypothesis**. A hypothesis may be **directional** or **non-directional** (i.e. the direction of a difference or relationship is or is not stated). A good hypothesis should be **operationalised** so that the variables are in a form that can be easily tested.

Investigator effects and other problems – Investigators may communicate their expectations unwittingly to participants (**investigator** or **experimenter effects**), thus leading participants to fulfil the investigator's expectations. Other problems include **demand characteristics** (features of an experiment that may cue participants to behave in predictable ways) and **social desirability bias** (participants wish to present themselves in a good light).

Pilot study – A small-scale trial run of a research study to test any aspects of the design, with a view to making improvements.

ISSUES OF RELIABILITY, VALIDITY AND SAMPLING

Reliability refers to how much we can depend on any particular measurement, for example, the measurement of a table, the measurement of a psychological characteristic such as IQ, or the findings of a research study. In particular we want to know whether, if we repeat exactly the same measurement/test/study we can be sure that we would get the same result. If not, our measurement is unreliable.

Validity is related to reliability because, if a measurement is not reliable (consistent), then a study cannot be valid i.e. it cannot be 'true' or legitimate. For example a researcher might measure intelligence using an intelligence test. If the same person is tested on several occasions using the same test and the results change each time then the intelligence test lacks reliability – and it also lacks validity because the scores are meaningless.

However a measurement may be reliable but still lack validity. For example, a person may take an IQ test and then take the same test several months later. Their score may be consistent, so it is a reliable test. However the items on the test may simply assess what a person learned at school rather than 'intelligence'. In which case the test is lacking validity (meaningfulness).

THE ISSUE OF RELIABILITY

We can consider reliability in relation to different research methods, including different *types* of reliability (internal and external), as well as looking at how reliability can be assessed and improved.

Experimental research

In the context of an experiment, reliability refers to the ability to repeat a study and obtain the same result i.e. **replication**. It is essential that all conditions are the same, otherwise any change in the result may be due to changed conditions.

Observational techniques

Observations should be consistent, which means that ideally two or more observers should produce the same record. The extent to which the observers agree is called inter-rater or **inter-observer reliability**, calculated by dividing total agreements by the total number of observations. A result of 0.80 or more suggests good inter-observer reliability.

The reliability of observations can be improved through training observers in the use of a coding system/behaviour checklist.

Self-report techniques

There are two different types of reliability which are particularly apparent when thinking of self-report techniques such as questionnaires and interviews.

- **Internal reliability** is a measure of the extent to which something is consistent within itself. For example, all the questions on an IQ test (which is a kind of questionnaire) should be measuring the same thing.
- **External reliability** is a measure of consistency over several different occasions. For example, if the same interview by the same interviewer with the same interviewee was conducted one day and then again a week later, the outcome should be the same, otherwise the interview is not reliable.

Reliability also concerns whether two interviewers produce the same outcome. This is called **inter-interviewer reliability**.

There are various ways to assess reliability such as using the **split-half method** to compare a person's performance on two halves of a questionnaire or test. If the test is assessing the same thing in all its questions then there should be a close correlation in the scores derived from both halves of the test, a measure of internal reliability. A second method of assessing reliability is the **test-retest method** where a person is given a questionnaire/interview/test on one occasion and then this is repeated again after a reasonable interval (e.g. a week or a month). If the measure is reliable the outcome should be the same every time.

THE ISSUE OF VALIDITY

All research strives to be high in validity. Any flaws must be minimised in order to draw valid conclusions from any study. There are two kinds of validity:

- **Internal validity** concerns what goes on inside a study – whether the researcher did test what he (or she) intended to test.
- **External validity** concerns things outside a study – the extent to which the results of the study can be generalised to other situations and people. The term **ecological validity** is often used as another term for external validity.

Experimental research

Internal validity is affected by extraneous variables (EVs) which may act as an alternative IV. Therefore changes in the DV are due to EVs rather than the IV, and conclusions about the effect of the IV on the DV are erroneous.

Many people think that all **laboratory experiments** are low in external validity, and that **field experiments**, conducted in more natural surroundings, are seen as high in external validity. This is not necessarily true. In some cases the contrived, artificial nature of the laboratory setting is not particularly relevant to the behaviour being observed (such as a memory task) and therefore it can be generalised to everyday situations (external validity). Also, in some cases, field experiments can be very contrived and artificial. And, as mentioned before, more control is possible in a laboratory.

It is often more important to consider issues such as whether the participants were aware they were being studied (which reduces the realism of their behaviour) and whether the task itself (rather than the setting) was artificial and thus low in **mundane realism**, which reduces the generalisability of the results.

Observational techniques

In terms of internal validity, observations will not be valid (nor reliable) if the coding system/behaviour checklist is flawed. For example, some observations may belong in more than one category, or some behaviours may not be codeable, which reduces the internal validity of the data collected.

The internal validity of observations is also affected by **observer bias** – what someone observes is influenced by their expectations. This reduces the objectivity of observations.

Observational studies are likely to have high ecological validity because they involve more natural behaviours – though, as we have seen, naturalistic research is not necessarily higher in ecological validity.

SAMPLING TECHNIQUES

The aim of all psychological research is to be able to make valid generalisations about behaviour. In any research project only a small number of participants are studied because larger groups involve more time and money. Psychologists use **sampling** techniques which will minimise cost while maximising generalisability.

Opportunity sample – Participants are selected by using those people who are most easily available. This is the easiest method to use but it is inevitably biased because the sample is drawn from a small part of the target population. For example, if you selected your sample from people walking around the centre of a town on a Monday morning, it would be unlikely to include professional people (because they are at work).

Volunteer sample – Participants are selected by asking for volunteers, for example placing an advertisement on a college noticeboard. This method can access a variety of participants if the advertisement is, for example, in a national newspaper, which would make the sample more representative. However such samples are inevitably biased because participants are likely to be highly motivated and/or with extra time on their hands (= **volunteer bias**).

Random sample – Participants are selected using a random number technique. First, all members of the target population are identified (e.g. all the members of one school or all the residents of a town) and then individuals are selected either by the lottery method (numbers drawn from a 'hat') or using a random number generator.

This method is potentially unbiased because all members of the target population have an equal chance of selection, though in the end a researcher may still end up with a biased sample because some people refuse to take part.

Stratified and quota samples – Sub-groups (strata) within a population are identified (e.g. boys and girls or different age groups). Then a predetermined number of participants is taken from each sub-group in proportion to their representation in the target population, e.g. if there are twice as many boys than girls in the target population then the sample will consist of twice as many boys. In stratified sampling this is done using random techniques, in quota sampling it is done using opportunity sampling.

This method is more representative than other methods because there is proportional representation of sub-groups. However selection within each sub-group may be biased, for example because of opportunity sampling.

Snowball sampling – In some studies it is difficult to identify suitable participants. For example, in a study of eating disorders, a useful technique is to start with one or two people with eating disorders and ask them to direct you to some other people with eating disorders and so on. This is useful when conducting research with participants who are not easy to identify, but is prone to bias because researchers may only contact people within a limited section of the population.

Self-report techniques

There are several ways to assess the internal validity of self-report techniques:

- **Face validity**: Does the test look as if it is measuring what the researcher intended to measure. For example, are the questions obviously related to the topic?
- **Concurrent validity**: This can be established by comparing performance on a new questionnaire or test with a previously established test on the same topic.

The external validity of self-report techniques is likely to be affected by the sampling strategies used which may create a biased sample.

*Sampling techniques are also used in observational studies as we have previously discussed – **time sampling** and **event sampling** (see page 280).*

*A **systematic sample** – e.g. taking every tenth person from a register – is often mistaken for a random sample. However, it is random if the first person is selected randomly!*

KEY CONCEPTS

The '**population**' (or target population) refers to all the people about whom a researcher wishes to make a statement. The aim is to select a **representative sample** from this target population – a small group who *represent* the target population in terms of characteristics such as age, IQ, social class, relevant experiences and so on.

The importance of representativeness is the ability to *generalise* from the sample to the target population. A sample that is not representative is described as **biased** i.e. leaning in one direction.

CAN YOU...? No.16.4

...1 Explain what is meant by validity and reliability, internal and external validity, and internal and external reliability.

...2 In each of the following studies describe **two** features of the study that might affect the validity of the data being collected and how the validity could be improved.

(a) A psychologist conducts interviews with mothers about their attitudes towards day care.
(b) A psychologist conducts a study to see if students do more homework in the winter or spring term. To do this he asks students to keep a diary of how much time they spend on homework each week.

...3 In each of the following studies suggest how reliability could be assessed.

(a) A psychologist intends to use a repeated measures design to test participants' memories in the morning and afternoon. He uses two tests of memory.
(b) A psychologist interviews teenage girls about their dieting.

...4 Some psychology students plan to conduct an observational study on the effects of different dress styles – to see if men look more at girls dressed casually or smartly.

(a) Identify **two** ways you could operationalise 'being dressed "casually"'.
(b) Identify **one** way in which you could ensure reliability among the different observers and explain how to do this.
(c) Describe what sampling technique you might use for making the observations.
(d) Explain **one** feature of the study that might affect the validity of the data being collected.

...5 On the previous spread **four** studies have been described. For each study identify an appropriate sampling technique and give **one** strength and **one** weakness of that technique in the context of the study.

...6 Why might a volunteer sample be preferable to an opportunity sample?

...7 Explain the difference between a random sample and a systematic sample.

ETHICAL CONSIDERATIONS IN PSYCHOLOGICAL RESEARCH

Any professional group, such as psychologists, doctors or solicitors, has a duty to behave in an ethical manner i.e. to behave with a proper regard for the rights and feelings of others. For psychologists this encompasses the treatment of patients, and the responsibilities of researchers towards their participants – human or non-human.

There is an important point to remember where ethical issues are concerned – there are no right or wrong answers because these are *issues* i.e. topics where there are conflicting points of view. Professional organisations such as the BPS (British Psychological Organisation) and APA (American Psychological Association) provide guidance for psychologists about how to behave. Such guidance is always being updated in order to keep up with changing viewpoints and new moral dilemmas (e.g. research on the internet).

WWW You can read the BPS code of conduct and other statements about ethical practice at www.bps.org.uk/the-society/code-of-conduct/code-of-conduct_home.cfm

SOCIALLY SENSITIVE RESEARCH

Despite the existence of ethical guidelines, some broader ethical issues arise in 'socially sensitive' areas. Sieber and Stanley (1988) defined **socially sensitive research** as '…studies in which there are potential social consequences or implications, either directly for the participants in research or the class of individuals represented by the research' (Sieber and Stanley, 1988).

One of the most controversial avenues of research has been consideration of inter-racial differences in IQ. Some evidence suggests that, in terms of IQ, black children may be innately inferior. Even though such research may be flawed (e.g. it ignores social conditions) such 'scientific' evidence can be used to support divisive and discriminatory social policies. Other areas of social sensitivity include research on drug abuse or sexual orientation.

The ethical question concerns whether or not such research should be conducted. If research is not conducted, such groups may miss out on any potential benefits from the research (e.g. increased funding or wider public understanding). Also, ignoring these important areas of research would amount to an abdication of the 'social responsibilities' of the psychological researcher (i.e. their duty to society to study important areas of human behaviour).

An understanding of the nature of socially-sensitive research should focus psychologists more clearly on the implications of their findings and the worrying potential, as Sieber and Stanley (1988) suggest, to on occasion offer, 'scientific credibility to the prevailing prejudice'.

▶ Concern for the protection of human participants in research has its roots in the Nuremberg Code (1947), a document designed to protect against atrocities such as those uncovered by the Nuremberg Trials following World War II. The Nuremberg Code was the first ethical code of practice. The APA produced the first code for psychologists in 1953 (a document of 170 pages).

ETHICAL ISSUES WITH HUMAN PARTICIPANTS

Ethical issues

The first main issue relates to **informed consent** and **deception**. Ideally, participants should be given the opportunity to know about all aspects of any research before agreeing to take part. This is a basic right stemming from the inhumane experiments conducted in concentration camps such as Auschwitz-Birkenau in the Second World War. However, the issue arises because full information may compromise the integrity of a study (e.g. knowing the full aims may alter participants' behaviour, rendering the results meaningless).

The second main issue relates to **harm**, and what constitutes too much harm. For example Ainsworth argued in her strange situation research with infants, that the distress they experienced was no greater than that experienced in everyday life (Ainsworth *et al.*, 1978).

In both cases the decision as to what is acceptable or not acceptable is open to debate.

Code of conduct

The current BPS code of ethics and conduct (BPS, 2006) identifies four ethical principles and includes advice on how these should be dealt with:

1 ***Respect*** for the dignity and worth of all persons. This includes standards of **privacy** and **confidentiality** and **informed consent**. Observations of behaviour in public without informed consent are only acceptable in situations where the people being studied would reasonably expect to be observed by strangers.

Intentional **deception** (lack of informed consent) is only acceptable when it is necessary to protect the integrity of research and when the nature of the deception is disclosed to participants at the earliest opportunity. One way to judge deception is to consider whether participants are likely to object or show unease when debriefed (see below), in which case the deception may be judged unacceptable.

Participants should be aware of the **right to withdraw** from the research at any time.

2 ***Competence*** – Psychologists should maintain high standards in their professional work.

3 ***Responsibility*** – Psychologists have a responsibility to their clients, to the general public and to the science of Psychology. This includes protecting participants from physical and psychological **harm** as well as **debriefing** at the conclusion of their participation to inform clients of the nature and conclusions of the research, to identify any unforeseen harm, and to arrange for assistance if needed.

4 ***Integrity*** – Psychologists should be honest and accurate. This includes reporting the findings of any research accurately and acknowledging any potential limitations. It also includes bringing instances of misconduct by other psychologists to the attention of the BPS.

Dealing with ethical issues

The code of conduct offers **ethical guidelines** for psychologists to follow. In conjunction with such guidelines psychologists deal with ethical issues by using **ethical committees** to assess research proposals, by punishing psychologists who contravene the code with disbarment from the society, and by educating students and qualified psychologists about their duties as researchers.

ETHICAL ISSUES WITH NON-HUMAN ANIMALS

Although the vast majority of investigations in psychology involve the study of *humans*, there are several reasons why psychologists may choose to carry out research using non-human animals:

- Animals may be studied simply because they are fascinating to study in their own right and such research may ultimately benefit animals.
- Animals offer the opportunity for greater control and objectivity in research procedures. Much of behaviourist theory was established using animal studies for just this reason, for example animals in the Skinner box (see page 124).
- We may use animals when we can't use humans. Animals have been exposed to various procedures and events that would simply not be possible with human beings. For example Harlow's research (1959) with rhesus monkeys and wire 'mothers' showed that contact comfort was a more essential requirement for primates than food.
- Human beings and non-human animals have enough of their physiology and evolutionary past in common to justify conclusions drawn from experiments involving one, to the other. However it can be argued that animals tested under stressful conditions may provide very little useful information.

Moral justification

The question still remains as to whether 'science at any cost' is justifiable.

Sentient beings – Do animals experience pain and emotions? In terms of pain there is evidence that they *respond* to pain but this may not be the same as conscious awareness. However, there is some evidence that animals other than primates have self-awareness (see page 128). In addition, some humans, such as brain-damaged individuals, lack sentience, but wouldn't be used in research without consent.

Speciesism – Peter Singer (1990) argued that discrimination on the basis of species is no different from racial or gender discrimination and thus suggests that the use of animals is an example of 'speciesism', similar to racism or sexism. However Gray (1991) argues that we have a special duty of care to humans, then speciesism is not equivalent to, for example, racism.

Animal rights – Singer's view is a utilitarian one, i.e. whatever produces the greater good for the greater number is ethically acceptable, so if animal research can alleviate pain and suffering it is justifiable. Tom Regan (1984), on the other hand, argues that there are no circumstances under which animal research is acceptable. Regan claims that animals have a right to be treated with respect and should *never* be used in research. The issue of 'animal rights' can be challenged by examining the concept of rights – having rights is dependent on having responsibilities in society, i.e. as citizens. It can therefore be said that as animals do not have responsibilities, they have no rights.

Existing constraints

Animal research is very strictly controlled. The BPS publishes guidelines for research with animals but, more importantly, there is legislation. In the UK, the Animals (Scientific Procedures) Act (1986) requires that animal research only takes place at licensed laboratories with licensed researchers on licensed projects. Such licences are only granted if:

- Potential results are important enough to justify the use of animals.
- The research cannot be done using non-animal methods.
- The minimum number of animals will be used.
- Any discomfort or suffering is kept to a minimum by appropriate use of anaesthetics or painkillers.

The 3 Rs were proposed by Russell and Birch (1959) – reduction (use fewer animals), replacement (where possible use alternative methods such as brain scans), and refinement (use improved techniques to reduce stress). The House of Lords (2002) endorsed the principle of the 3Rs with respect to animal research. Nevertheless, in the UK, the need for animal research continues. For example British law requires that any new drug (such as antidepressants) must be tested on at least two different species of live mammal.

▼ When considering the use of non-human animals in psychological research, it is important to put your emotions to one side and consider the logic of the arguments. It is also important to remember that much *psychological* research with animals does not involve physical interventions – though psychological treatments may be equally damaging.

WWW If you wish to read more about animal research, the BBC website presents some excellent information – see www.bbc.co.uk/ethics/animals/

CAN YOU...? No.16.5

...1 For each of the following studies, identify **one** ethical issue that might arise, and suggest how the researcher might deal with it.

(a) A correlation of pupil IQ scores and GCSE results.

(b) Interviewing teenage girls about their dieting habits.

(c) An observational study of the way in which children cross the road going to and from school.

(d) A psychologist decides to conduct a field experiment to see whether people are more likely to obey someone in a uniform or dressed in a casual suit.

(e) A school decides to conduct a natural experiment to see if the students doing a new maths programme do better in their GCSE maths exam than a group of students using the traditional learning methods.

(f) An experiment to test the effect of self-esteem on performance. Participants are given a self-esteem questionnaire and then given a false score (told they either have high or low self-esteem).

(g) A teacher asks her students to take part in a research project, telling them it is about eating habits whereas it is really about eating disorders.

...2 Describe **one** example of socially sensitive research and consider the pros and cons of conducting such research.

...3 What protection exists to ensure that animals involved in psychological research are treated ethically?

...4 Identify **two** examples of psychological research that have used non-human animals and discuss the ethical issues raised by this study.

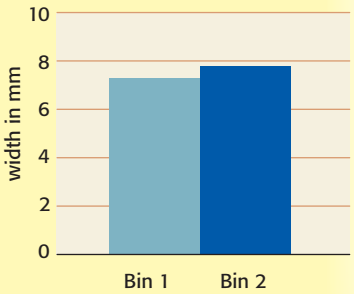
INFERENTIAL ANALYSIS, PROBABILITY AND SIGNIFICANCE

You may have heard the phrase 'statistical tests' – for example in a newspaper report that claims 'statistical tests show that women are better at reading maps than men'. If we wanted to know if women are better at reading maps than men we could not possibly test all the women and men in the world, so we just test a small group of women and a small group of men. If we find that the sample of women are indeed better with maps than the sample of men, then we *infer* that the same is true for all women and men. However, it isn't quite as simple as that because we can only make such inferences using statistical (or inferential) tests. Such statistical tests are based on probabilities, so we will start the topic of inferential analysis by looking at probability.

▲ We all have an intuitive sense of probability. Would you like to buy a £50 raffle ticket to win this car? If the promoter is only selling 10 tickets your answer would probably be 'yes' because your chances of winning would be fairly high, whereas, if he is selling 1000 tickets, you might think again. Probability is about chance – how probable is it that you would win if 1000 tickets are sold? You have a 1 in 1000 chance of winning or 0.1% probability (one divided by ten = 0.1%).

*In the AS Psychology course you looked at **descriptive statistics** such as measures of central tendency (mean, median and mode), measures of dispersion (standard deviation and range), and methods of graphical representation (scattergrams and bar charts). We can draw conclusions from descriptive statistics, such as using a graph to see that one group of participants did better than another group. But we don't know whether this difference is **significant**. Inferential statistics allows us to determine whether a difference is significant. Remember that significance is the extent to which something is particularly unusual.*

▼ Graph showing the mean width for both bins



PROBABILITY

Inferential statistics allow psychologists to draw conclusions based on the probability that a particular pattern of results could have arisen by chance. If it could have arisen by chance, then it would not be correct to conclude, in the example above, that women are better than men at reading maps, or in the case of a correlational study that there was a real association between two variables. However, if it could *not* have arisen by chance, or if it is extremely unlikely to have arisen by chance, then the pattern is described as **significant**.

SIGNIFICANCE

Consider the following example from the statistician Hugh Coolican (2004): 'At my local chippy I am convinced that they save money by giving some people rather thin chips (meaning they can get more chips from each potato). There are two chip bins under the counter – they claim they are the same but I suspect they are different. So I (sadly) tried an experiment. I asked for one bag of chips from bin 1 and one lot from bin 2, and I went home and measured the width of the chips in each bag.'

In order to use an inferential test we need a **null hypothesis** and an **alternative hypothesis**.

- The null hypothesis is: 'The two bins contain chips of an equal average width'.
- The alternative hypothesis is: 'One bin has thinner chips on average than the other'.

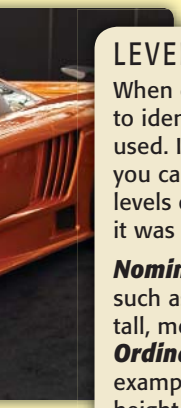
In fact there was a very small difference between the average width of the chips in each bag (as you can see in the bar chart on the left), but nothing of any significance. We would expect small differences between samples (bags of chips) just because things do vary a little each time you do them – this is simply random variation or 'chance'. What we are looking for is a sufficiently large (significant) difference between samples to be sure that the bins *are* actually different. Otherwise we assume the bins (populations) are the same (accept the null hypothesis).

CHANCE

In order to work out whether a difference is or is not significant we use inferential tests. Such tests are based on some cunning maths that you don't need to know about. They permit you to work out, for a given probability, whether a pattern in the data from a study could have arisen by **chance** or whether the effect occurred because there is a real difference/correlation in the populations from which the samples were drawn.

But what do we mean by 'chance'? We simply decide on a probability that we will 'risk'. You can't be certain that an observed effect was not due to chance but you can state *how* certain you are. In general, psychologists use a **probability** of $p \leq 0.05$, which means that there is a 5% possibility that the results did occur by chance. In other words a 5% probability that the results occurred even if there was no *real* difference/association between the populations from which the samples were drawn. In our example about the chip bins (left), the samples differed slightly but this might be (a) due to chance, or (b) because the samples did come from two different populations (large chip bin and small chip bin). Ultimately we are interested in making a statement about the population(s) from which the samples are drawn.

In some studies psychologists want to be more certain – such as when they are conducting a replication of a previous study or considering the effects of a new drug on health. Then, researchers use a more stringent probability, such as $p \leq 0.01$ or even $p \leq 0.001$. In other studies a more lenient level of $p \leq 0.10$ might be used, such as when conducting research into a new topic. This chosen value of 'p' is called the **significance level**.



LEVELS OF MEASUREMENT

When deciding which test to use you may need to identify the **level of measurement** that was used. If the data are, for example, nominal then you cannot use Spearman's test. We described the levels of measurement in the AS book but thought it was worth repeating again!

Nominal – The data are in separate categories, such as grouping your class into people who are tall, medium or short.
Ordinal – Data are ordered in some way, for example lining up your classmates in order of height. The 'difference' between each item is not the same.
Interval – Data are measured using units of equal intervals, such as when counting correct answers or measuring your classmates' heights.
Ratio – There is a true zero point as in most measures of physical quantities.
NOIR – An acronym to help remember the four levels of measurement of data: *nominal*, *ordinal*, *interval* and *ratio*.

TYPE 1 AND TYPE 2 ERRORS

In general, psychologists use a 5% significance level. One reason for this is that a 5% degree of uncertainty is usually acceptable if it is not a life and death matter, whereas, for example, research on the side-effects of drugs is likely to select a 1% significance level because we would want to be very careful about taking chances.

There is another reason for using the 5% level. If you use a level of significance that is too high (lenient), such as 10%, then you may reject a null hypothesis that is true. Consider this example – imagine I take a pregnancy test and it is positive, leading me to accept the hypothesis that I am pregnant (thus rejecting the null hypothesis that I am not pregnant). What if the test was wrong, so I had a false positive? This is called a **Type 1 error** – rejecting a null hypothesis that is true. In research the likelihood of a Type 1 error is increased if the significance level is too high (e.g. 10%).

On the other hand the result of my pregnancy test could be negative even though I was actually pregnant, thus I erroneously accept the null hypothesis that I am not pregnant. This is a **Type 2 error** – accepting a null hypothesis that is in fact not true. In research, the likelihood of a Type 2 error is increased if the significance level is too low (stringent), such as 1%.

	In truth, the null hypothesis is...	
	TRUE	***FALSE***
Reject null hypothesis	**Type 1 error**	**Correct!**
Accept null hypothesis	**Correct!**	**Type 2 error**

USING INFERENTIAL STATISTICAL TESTS

Different tests

Different inferential tests are used for different research designs. For example, if a study involves looking at the correlation between two variables, then the test used to determine whether or not there is a significant correlation is a correlational test such as Spearman's *rho*. There are four tests that you are required to study and these are each considered on the next four spreads – **Spearman's *rho*, chi-square, Mann-Whitney *U*** and **Wilcoxon T.**

Observed and critical values

Each inferential test involves taking the data collected in a study and doing some arithmetical calculations which produce a single number called the **test statistic**. In the case of Spearman's correlation test, that test statistic is called *rho*, whereas for the Mann-Whitney test it is *U*. The *rho* value calculated for any set of data is called the **observed value** (because it is based on the observations made). To decide if the observed value is significant, this figure is compared to another number, found in a **table of critical values**; this is called the **critical value**, and is the value that a test statistic must reach in order for the null hypothesis to be rejected. There are different tables of critical values for each different inferential test (see, for example, page 289). To find the appropriate critical value in a table you need to know:

1 **Degrees of freedom** (df) – In most cases you get this value by looking at the number of participants in the study (N). In studies using an independent groups design there are two values for N (one for each group of participants), called N_1 and N_2. In the case of the chi-square test you calculate *df* on the basis of how many 'cells' there are.
2 **One-tailed or two-tailed test** – If the hypothesis was a directional hypothesis, then you use a one-tailed test, if it was non-directional you use a two-tailed test.
3 **Significance level** selected, usually $p \leq 0.05$.

The importance of R

Some tests are significant when the observed value is equal to or exceeds the critical value, for others it is the reverse (the size of the difference between the two is irrelevant). You need to know which, and you will find it is stated underneath each table. One way to remember is to see if there is a letter R in the name of the test. If there is an R then the observed value should be g**R**eate**R** than the critical value (e.g. for Spearman's and chi-square). If there is no R (e.g. Mann-Whitney and Wilcoxon) then the observed value should be less than the critical value.

*You should remember **directional** and **non-directional** hypotheses from your AS studies. For example 'Women are better at reading maps than men' is a directional hypothesis because we have predicted in which direction the results will occur (women do better). 'Women and men are different in their map-reading abilities' would be a non-directional hypothesis because it does not predict which group will do better.*

*Directional hypotheses are also called '**one-tailed**' because they are only concerned with one possibility (or 'tail') – our directional hypothesis ignores the other possibility that men might be better than women. The non-directional hypothesis covers both possibilities and thus is '**two-tailed**'.*

CAN YOU...? No.16.6

...**1** Identify the letter used to signify significance level.
...**2** Give **two** examples of descriptive statistics.
...**3** Explain the reason for using statistical (inferential) tests.
...**4** Explain what is meant by the phrase 'significant at $p \leq 0.05$'.
...**5** Suggest why a researcher may choose to use $p \leq 0.01$ in preference to $p \leq 0.05$. (Try to give **two** reasons.)
...**6** Give the general name for the value that is worked out using an inferential test.
...**7** Give the general name given to the number, found in a significance table, that is used to judge the observed value produced by an inferential test.
...**8** Identify the **three** pieces of information used to find this value.
...**9** Identify the level of measurement that would be used in the examples below:

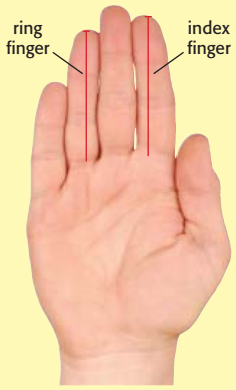
(a) Rating how stressful certain experiences are.
(b) Counting the days a person has had off school.
(c) Asking people to indicate the reasons for days off school.

...**10** Think of a way to remember which is a Type 1 error and which is a Type 2 error.

INFERENTIAL TESTS: SPEARMAN'S *RHO*

The first inferential test we will look at is a test of correlation – Spearman's *rho*. It is used to determine whether the correlation between two co-variables is significant or not. For example, the study of stress and illness by Rahe *et al.* (1970, see page 281) found a correlation of + .118 between the number of times a participant was ill and their stress score as measured by the SRRS (social readjustment rating scale). A figure of zero would be no correlation, whereas a figure of +1.0 would be a perfect positive correlation. A correlation of +.118 may sound like a rather insignificant correlation but in fact it is significant.

The observed value of +.118 was calculated using an inferential statistical test such as Spearman's *rho*. This observed value is then compared to the critical value found in a table of critical values, such as the table on the far right, to see whether the observed value is significant. In this study the number of participants was over 2700 and therefore .118 was significant. (As the number of participants increases, the value needed for significance decreases.) Incidentally if the observed value had been –.118 this would still be significant – a significant *negative* correlation.



▲ **Literacy hand**
People with short ring fingers and long index fingers are better at literacy.

▲ **Numeracy hand**
People with long ring fingers and short index fingers are more likely to excel in numeracy.

HEALTH WARNING – ETHICS

If you do conduct your own research studies, make sure no participants are younger than 16 and that you seek fully informed consent. If using sensitive information, such as tests of maths ability, then you must protect participants' identities.

FINGER LENGTH AND EXAM PERFORMANCE

A number of studies have looked at the relationship between finger length and various abilities such as numeracy or literacy. For example a recent study by Brosnan (2008) examined finger length in 75 British children aged between six and seven (boys and girls), and found that children with a higher digit ratio between their index and ring fingers were more likely to have a talent in maths, while those with a shorter digit ratio were more likely to have a talent in literacy.

This relationship is thought to be due to biological factors, specifically the production of **testosterone** and **oestrogen** in the brain. Male babies are exposed to more testosterone (a male hormone) during prenatal development and this affects their finger length. Testosterone also promotes the development of the areas of the brain which are often associated with spatial and mathematical skills, whereas oestrogen (a female hormone) is thought to do the same in the areas of the brain which are often associated with verbal ability.

In order to study the correlation between finger length and numeracy/literacy the researchers took photocopies of both the right and left hands of the children and measured the length of the index and ring fingers. They divided the length of the index finger by the length of the ring finger to calculate each child's 'digit ratio'.

The digit ratios were then correlated with the results from their National Standard Assessment Tests (SATs) for numeracy and literacy.

DO-IT-YOURSELF

You can repeat (replicate) this study using GCSE scores instead of SATs results, or you could use online tests of literacy and/or numeracy. In order to determine if your results are significant, follow the worked example on the facing page.

CAN YOU…? No.16.7

…1 Identify **two** ethical problems that might arise when conducting a study on finger length and numeracy, and state how these could be dealt with.

…2 Suggest problems that might occur when dealing with the ethical problems in the manner you suggested

…3 Identify the co-variables in the study by Brosnan (above).

…4 Identify the intervening variable in this study (the variable that links finger length to, for example, numeracy).

…5 If you were going to study the relationship between finger ratio and literacy, state a possible alternative and null hypothesis for this study.

…6 Draw a scattergram of the results on the right to check the outcome of the inferential test – does your graph show the same relationship as found by the inferential test?

…7 If you conduct a study yourself, produce a report of the study using the normal conventions (see 'Conventions for reporting psychological investigations' on page 278). Remember to use a scattergram to 'eyeball' your data as well as conducting an inferential test.

MORE DO-IT-YOURSELF IDEAS FOR STUDIES USING A CORRELATIONAL ANALYSIS

- **Stressful life events and illness**. In your AS studies you looked at the relationship between stressful life events or daily hassles and illness. You can replicate such a study and check whether your correlations were significant.
- **Reaction time and number of hours of sleep**. Does sleep deprivation have any effects? For example it might be related to poor reaction time. You can measure reaction time using an online test.
- **Reaction time and time spent playing computer games**. Perhaps playing computer games is related to reaction time.
- **Working memory and IQ** are predicted to be positively correlated. You can find tests for both at www.bbc.co.uk.

WHEN TO USE SPEARMAN'S RANK CORRELATION (RHO) TEST

- The hypothesis predicts a **correlation** between two co-variables.
- The two sets of data are pairs of scores from one person or thing = i.e. they are **related**.
- The data are **ordinal** or **interval** (i.e. not nominal). See page 000 for an explanation.

SPEARMAN'S RANK CORRELATION TEST – A WORKED EXAMPLE

STEP 1. State the alternative and null hypothesis

Alternative hypothesis: The digit ratio between index finger and ring finger is positively correlated to numeracy skills. (This is a directional hypothesis, therefore requiring a one-tailed test.)

Null hypothesis: There is no correlation between digit ratio and numeracy skills.

STEP 2. Record the data, rank each co-variable, and calculate the difference

- Rank A and B separately, from low to high (i.e. the lowest number receives the rank of 1).
- If there are two or more of the same number (tied ranks), calculate the rank by working out the mean of the ranks that would have been given.

You will never be required to do any calculations in an exam. The reason it is a good idea to have a go is that it gives you a better 'feel' for the test and also helps you understand how to deal with observed and critical values, and draw a conclusion. You can, however, avoid the actual calculation by using an electronic version – look online or use one in Excel (Microsoft Office).

Participant number	Digit ratio	Numeracy score	Rank A	Rank B	Difference between rank A and rank B (d)	d2
1	1.026	8	10	2.5	7.5	56.25
2	1.000	16	5.5	9	–3.5	12.25
3	1.021	10	9	5	4.0	16.0
4	0.991	9	4	4	0	0
5	0.984	15	3	8	–5.0	25.0
6	0.975	14	1	7	–6.0	36.0
7	1.013	12	7	6	1	1.0
8	1.018	8	8	2.5	5.5	30.25
9	0.982	17	2	10	–8.0	64.0
10	1.000	5	5.5	1	4.5	20.25
N = 10				Σd^2 (sum of differences squared) = 261.0		

STEP 3. Find observed value of rho *(the correlation coefficient)*

$$rho = 1 - \frac{6\Sigma d^2}{N(N^2-1)} = 1 - \frac{6 \times 261.0}{10 \times (100-1)} = 1 - \frac{1566}{990} = 1 - 1.58 = -0.58$$

STEP 4. Find the critical value of rho

N=10, the hypothesis is directional, therefore a one-tailed test is used.

Look up critical value in table of critical values (on right).

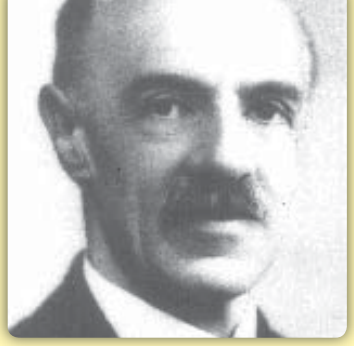
For a one-tailed test where N=10, the critical value of *rho* ($p \leq 0.05$) = 0.564

Note that the observed value is negative – when comparing this figure to the critical value, only the value, not the sign, is important. The sign does, however, tell you whether the correlation is positive or negative. If the prediction was one-tailed and the sign (and therefore the correlation) is not as predicted, then the null hypothesis must be retained.

STEP 5. State the conclusion

As the observed value (0.58) is greater than the critical value (0.564) we could reject the null hypothesis (at $p \leq 0.05$) and therefore could conclude that digit ratio is correlated with numeracy.

Note, however, that in this case the sign is in the wrong direction – a positive correlation was predicted but a negative correlation was found. This means that in fact we have to accept the null hypothesis. If we had predicted a negative correlation then we could have rejected the null hypothesis – though you should also remember that there is a 5% chance that we are wrong because p≤0.05.



▲ Charles Edward Spearman (1863–1945)

ALTERNATIVE AND NULL HYPOTHESIS

On the previous spread we introduced the **alternative** and **null hypothesis**. The term 'alternative hypothesis' (H_1) is used because it is the alternative to the null hypothesis (H_0). The null hypothesis is required because inferential tests are looking at whether our samples come from a population where there is no relationship (in which case the null hypothesis is true i.e. any relationship is due to chance) or whether our samples come from a population where there is a relationship (in which case we can reject the null hypothesis and accept the alternative).

The null hypothesis is a statement of no relationship (in a correlational analysis) or no difference. So it should always begin: 'There is no correlation between …' or: 'There is no difference between …'.

N =	One-tailed test	Two-tailed test
4	1.000	
5	0.900	1.000
6	0.829	0.886
7	0.714	0.786
8	0.643	0.738
9	0.600	0.700
10	0.564	0.648
11	0.536	0.618
12	0.503	0.587
13	0.484	0.560
14	0.464	0.538
15	0.443	0.521
16	0.429	0.503
17	0.414	0.485
18	0.401	0.472
19	0.391	0.460
20	0.380	0.447
21	0.370	0.435
22	0.361	0.425
23	0.353	0.415
24	0.344	0.406
25	0.337	0.398
26	0.331	0.390
27	0.324	0.382
28	0.317	0.375
29	0.312	0.368
30	0.306	0.362

◀ **Table of critical values of rho at 5% level ($p \leq 0.05$)**

Observed value of *rho* must be EQUAL TO or GREATER THAN the critical value in this table for significance to be shown.

Source: J.H. Zhar (1972). Significance testing of the Spearman rank correlation coefficient. *Journal of the American Statistical Association*, *67*, 578–80. (Reproduced with kind permission of the publisher.)

INFERENTIAL TESTS: CHI-SQUARE (χ^2) TEST

The second inferential test we will look at deals with **nominal data** i.e. data that is in categories. We use this test when we have counted how many occurrences there are in each category – called 'frequency data'. For example, we might be interested to find out whether men and women do actually differ in terms of their finger-length ratio. Research has found that adult women usually have ratios of one i.e. their index and ring fingers are of equal length. The average for men is lower, at 0.98, since they tend to have longer ring fingers than index fingers, suggesting greater exposure to testosterone in the womb. Of course the chi-square analysis (see below) does not *prove* this but can support this gender difference.

'Chi' is one of the letters of the Greek alphabet (pronounced as 'kie' to rhyme with 'pie'). The Greek symbol for chi is χ, which is why this symbol is used for the statistic for the chi-square test.

CHI-SQUARE TEST – A WORKED EXAMPLE FOR A 2 × 2 TABLE

Chi-square can be used to investigate a difference (as in the worked example on this page) or an association (as on the facing page).

STEP 1. State the alternative and null hypothesis

Alternative hypothesis: There is difference between men and women in terms of digit ratio (the ratio between the index and ring fingers). (This is a non-directional hypothesis, and therefore requires a two-tailed test.)

Null hypothesis: There is no difference between men and women in terms of digit ratio.

There are online programmes that will calculate chi-square for you, see for example http://math.hws.edu/javamath/ryan/ChiSquare.html (scroll about half way down the page). You can also use Excel.

STEP 2. Draw up a contingency table

	Male	Female	Totals
Digit ratio ≥1.00	5 (cell A)	12 (cell B)	17
Digit ratio <1.00	10 (cell C)	9 (cell D)	19
Totals	15	21	36

This is a 2 × 2 contingency table as there are two rows and two columns. On the facing page there is a 3 × 2 contingency table as there are three rows and two columns. The first number is always rows and the second number is columns (rows then columns, RC as in Roman Catholic).

STEP 3. Find observed value by comparing observed and expected frequencies for each cell.

The expected frequencies are calculated by working out how the data would be distributed across all cells in the table if there were no differences or pattern.

In some books Yates's correction is recommended but Coolican (1996) said that this was no longer current practice.

Many students get confused about the expected frequencies. These are not what the researcher expects – they are the frequencies that would occur if the data were distributed evenly across the table in proportion to the row and column totals.

	row × column / total = expected frequency (E)	Subtract expected value from observed value, ignoring signs (O–E)	Square previous value $(O–E)^2$	Divide previous value by expected value $(O–E)^2/E$
Cell A	17 × 15 / 36 = 7.08	5 – 7.08 = 2.08	4.3264	0.6110
Cell B	17 × 21 / 36 = 9.92	12 – 9.92 = 2.08	4.3264	0.4361
Cell C	19 × 15 / 36 = 7.92	10 – 7.92 = 2.08	4.3264	0.5463
Cell D	19 × 21 / 36 = 11.08	9 – 11.08 = 2.08	4.3264	0.3905

STEP 4. Add all the values in the final column

This gives you the observed value of chi-square as 1.984

STEP 5. Find the critical value of chi-square

Calculate degrees of freedom (*df*) by multiplying (rows – 1) × (columns – 1) = 1
Look up value in table of critical values (on the right).
For a two-tailed test, $df = 1$, the critical value of χ^2 ($p \leq 0.05$) = 3.84

▼ Table of critical values of chi-square (χ^2) ($p \leq 0.05$)

df	One-tailed test	Two-tailed test
1	2.71	3.84
2	4.60	5.99
3	6.25	7.82
4	7.78	9.49
5	9.24	11.07

Observed value of χ^2 must be EQUAL TO or GREATER THAN the critical value in this table for significance to be shown.

Source: abridged from R.A. Fisher and F. Yates (1974). *Statistical tables for biological, agricultural and medical research* (sixth edition). Longman.

The table shows one- and two-tailed values but statisticians argue that you can only test one-tailed (directional) hypotheses with a chi-square test.

STEP 6. State the conclusion

As the observed value (1.984) is less than the critical value (3.84) we must accept the null hypothesis (at $p \leq 0.05$) and therefore we conclude that there is no difference between men and women in terms of digit ratio.

WHEN TO USE THE CHI-SQUARE (χ^2) TEST

- The hypothesis predicts a **difference** between two conditions or an **association** between co-variables.
- The sets of data must be **independent** (no individual should have a score in more than one 'cell').
- The data are in **frequencies** (i.e. **nominal**). Frequencies must not be percentages.

Note – This test is unreliable when the *expected* frequencies (i.e. the ones you calculate) fall below 5 in any cell i.e. you need at least 20 participants for a 2 × 2 contingency table.

PARENTAL STYLE AND SELF-ESTEEM

Psychological research has identified three different parenting styles: authoritarian (parents dictate how children should behave), democratic (parents discuss standards with their children) and laissez-faire (parents encourage children to set their own rules). Buri (1991) found that children who experienced authoritarian parenting were more likely to develop high self-esteem.

DO-IT-YOURSELF

You can access the Parental Authority Questionnaire (PAQ) at http://faculty.sjcny.edu/~treboux/documents/parental%20authority%20questionnaire.pdf

There are various self-esteem questionnaires on the Internet.

CHI-SQUARE TEST – A WORKED EXAMPLE FOR A 3 × 2 TABLE

STEP 1. State the alternative and null hypothesis

Alternative hypothesis: Certain parental styles are associated with higher self-esteem in adolescence. (This is a non-directional hypothesis and therefore requires a two-tailed test.)

Null hypothesis: There is no association between parental style and self-esteem in adolescence.

STEP 2. Draw up a contingency table

In this case it will be 3 by 2 (rows first then columns)

	Self-esteem		
Parental style	High	Low	Totals
Authoritarian	10 (cell A)	4 (cell B)	14
Democratic	5 (cell C)	7 (cell D)	12
Laissez-faire	8 (cell E)	2 (cell F)	10
Totals	23	13	36

STEP 3. Find observed value by comparing observed and expected frequencies

	row × column / total = expected frequency	Subtract expected value from observed value, ignoring signs. $\|(O-E)\|$	Square the previous value $(O-E)^2$	Divide previous value by the expected value $(O-E)^2/E$
Cell A	14 × 23 / 36 = 8.94	10 – 8.94 = 1.06	1.1236	0.1257
Cell B	14 × 13 / 36 = 5.06	4 – 5.06 = -1.06	1.1236	0.2221
Cell C	12 × 23 / 36 = 7.67	5 – 7.67 = -2.67	7.1289	0.9294
Cell D	12 × 13 / 36 = 4.33	7 – 4.33 = 2.67	7.1289	1.6464
Cell E	10 × 23 / 36 = 6.39	8 – 6.39 = 1.61	2.5921	0.4056
Cell F	10 × 13 / 36 = 3.61	2 – 3.61 = -1.61	2.5921	0.7180

STEP 4. Add all the values in the final column

This gives you the observed value of chi-square (χ^2) = 4.0472

STEP 5. Find the critical value of chi-square (χ^2)

Calculate degrees of freedom (*df*) calculate (rows – 1) × (columns – 1) = 2
Look up critical value in table of critical values (on facing page).
For a two-tailed test, *df* = 2, the critical value of χ^2 ($p \leq 0.05$) = 5.99

STEP 6. State the conclusion

As the observed value (4.0472) is less than the critical value (5.99) we must accept the null hypothesis (at $p \leq 0.05$) and therefore conclude that there is no association between parental style and self-esteem in adolescence.

MORE DO-IT-YOURSELF IDEAS FOR STUDIES USING A CHI-SQUARE TEST

- **Gender and conformity**. Are women more conformist than men? Some studies have found this to be true though Eagly and Carli (1981) suggest that this is only the case for male-oriented tasks. Try different types of conformity tasks and see whether they have higher or lower levels of female conformity, for example ask questions on a general knowledge test which are related to male or female interests. The answers from previous 'participants' should be shown so you can see if your real participant conforms to the majority answer.
- **Sleep and age**. Research suggests that people sleep less as they get older (see page 10). Compare older and younger participants in terms of average numbers of hours of sleep.

CAN YOU...? No.16.8

...1 Draw a contingency table to show the following data – old and young participants are asked whether they sleep more or less than eight hours per night on average. Of the older people, 11 said they sleep more and 25 said they sleep less. Of the younger participants, 31 said they sleep more than eight hours and 33 said they sleep less.

...2 State an appropriate one-tailed alternative hypothesis and null hypothesis for this investigation.

...3 The observed value of chi-square for the data from question 1 is 3.02 (one-tailed test). Is this value significant? Explain your decision and state whether this means you can reject the null hypothesis.

INFERENTIAL TESTS: MANN-WHITNEY *U* TEST

The final two inferential tests required in the specification are called 'tests of difference'. They enable us to consider whether two samples of data are different or not different from each other i.e. that they are drawn from the same population (the null hypothesis) or from two different populations. Tests of difference are generally used for experiments. For example we might conduct an experiment to see if noisy conditions reduce the effectiveness of revision.

Case A – we could have two groups of participants:
- Group 1: participants revise in a silent room and are tested.
- Group 2: participants revise in a noisy room and are tested.

Case B – we might have two conditions:
- Condition 1: participants revise in a silent room and are tested.
- Condition 2: the same participants revise in a noisy room and are tested.

Case A is an **independent groups design** (we have two separate groups of participants). Case B (two conditions) is a **repeated measures design** as the same participants are tested twice.

The Mann-Whitney test is used for independent groups designs and the Wilcoxon test (on the next spread) is used for repeated measures designs.

In some experiments there are more than two conditions or groups – for example the study by Loftus and Palmer on leading questions had five different groups according to which verb was in the sentence (smashed, hit, etc). There are other inferential tests that are used for designs with more than two conditions/groups.

◀ The Capilano Suspension Bridge was used in the study by Dutton and Aron (see right). The bridge is narrow and long and has many arousal-inducing features: a tendency to tilt, sway, and wobble, creating the impression that one is about to fall over the side; very low handrails of wire cable contribute to this impression as well as a 230-foot drop to rocks and shallow rapids below.

FALLING IN LOVE

Psychologists have sought to explain the process of falling in love. One suggestion is that love is basically physiological arousal – arousal of your **sympathetic nervous system** which occurs when you are feeling scared or stressed or find someone physically attractive. Hatfield and Walster (1981) suggested that love is simply a label that we place on physiological arousal when it occurs in the presence of an appropriate object. A man or woman who meets a potential partner after an exciting football game is more likely to fall in love than he or she would be on a routine day. Likewise, a man or woman is more likely to fall in love when having experienced some bitter disappointment. The reason, in both cases, is to do with the two components of love: arousal and label.

This has been supported by various experiments, such as a memorable study by Dutton and Aron (1974). A female research assistant (unaware of the study's aims) interviewed males, explaining that she was doing a project for her psychology class on the effects of attractive scenery on creative expression. The interviews took place on a high suspension bridge (high arousal group, see left) or a narrow bridge over a small stream (low arousal). When the interview was over, the confederate gave the men her phone number and asked them to call her if they had any questions about the survey. Over 60% of the men in the high arousal condition did phone her, compared with 30% from the low arousal group, suggesting that the men had mislabelled their fear-related arousal as sexual arousal.

DO-IT-YOURSELF

Another study which investigated the two-factor theory of love (which might be easier to replicate) was conducted by White *et al.* (1981). In this experiment, high and low arousal were created by asking men to run on the spot for 2 minutes or 15 seconds respectively, and then showing a short video of a young woman. The more highly aroused men rated the woman as more attractive.

CAN YOU...? No.16.9

...1 Use descriptive statistics to summarise the results given in the worked example on the facing page, e.g. calculate measures of central tendency and dispersion, and also sketch an appropriate graph.

...2 A psychology class decides to replicate the study by White *et al.* on the right. Write an appropriate alternative and null hypothesis for this study.

...3 Is your alternative hypothesis directional or non-directional?

...4 The students check the significance of their results using the Mann-Whitney test and find that $U=40$ (there were 9 participants in one group and 13 in a second group). State what conclusion they could draw from their results.

...5 If you were going to write a report of this study, outline what would be included in each section of it (see 'Conventions for reporting psychological investigations' on page 278).

MORE DO-IT-YOURSELF IDEAS FOR STUDIES USING AN INDEPENDENT GROUPS TEST

- **Digit ratio and gender**. You can collect data on the digit ratios of men and women and analyse it using the Mann-Whitney test by comparing the scores for men and women.
- **The power of touch**. A number of studies have shown that people are more willing to comply with a request if you touch them lightly on their arm. For example Brockner *et al.* (1982) arranged for a confederate to approach participants as they left a phone box and ask for some money which had been left behind in the phone box. If the participant was touched lightly on the arm, 98% gave the money back, compared to 63% in the no-touch group.

*There are **three** kinds of experimental design – repeated measures, independent groups and finally, **matched pairs**. In a matched pairs study there are two groups of participants (as in independent groups design), however the groups are not independent, they are matched (e.g. on characteristics such as IQ, age, etc.). Therefore matched pairs experiments use repeated measures tests.*

THE MANN-WHITNEY *U* TEST – A WORKED EXAMPLE

STEP 1. State the alternative and null hypothesis

Alternative hypothesis: Male participants interviewed on a high bridge give higher ratings of the attractiveness of a female interviewer than those interviewed on a low bridge. (This is a directional hypothesis and therefore requires a one-tailed test).

Null hypothesis: There is no difference in the ratings of attractiveness given by those interviewed on a high or low bridge.

STEP 2. Record the data in a table and allocate points

- To allocate points, consider each score one at a time.
- Compare this score (the target) with all the scores in the other group.
- Give 1 point for every score that is higher than the target score.
- Give ½ point for every equal score.

STEP 3. Find observed value of U

U is the lower total number of points. In this case it is 16.5

STEP 4. Find the critical value of U

N_1 = number of participants in group 1
N_2 = number of participants in group 2
Look up the critical value in a table of critical values (below).
For a one-tailed test, $N_1 = 0$ and $N_2 = 14$, the critical value of U ($p \leq 0.05$) = 41

Note – when you have a one-tailed hypothesis, remember to check whether the difference is in the direction that you predicted. If it is not, you cannot reject the null hypothesis.

STEP 5. State the conclusion

As the observed value (16.5) is less than the critical value (41), we can reject the null hypothesis (at $p \leq 0.05$) and therefore conclude that participants interviewed on a high bridge give higher ratings of attractiveness to a female interviewer than those interviewed on a low bridge.

WHEN TO USE THE MANN-WHITNEY *U* TEST

- The hypothesis predicts a **difference** between two sets of data.
- The two sets of data are from separate groups of participants = **independent groups**.
- The data are **ordinal** or **interval** (i.e. not nominal). See page 287 for an explanation.

The Mann-Whitney test is named after the Austrian-born US mathematician Henry Berthold Mann and the US statistician Donald Ransom Whitney who published the test in 1947. They adapted a test designed by Wilcoxon which was for equal sample sizes.

Attractiveness ratings given by high bridge group	Points	Attractiveness ratings given by low bridge group	Points
7	1.5	4	10.0
10	0	6	8.5
8	1.0	2	10.0
6	3.5	5	9.5
5	7.0	3	10.0
8	1.0	5	9.5
9	0.5	6	8.5
7	1.5	4	10.0
10	0	5	9.5
9	0.5	7	7.0
		9	3.0
		3	10.0
		5	9.5
		6	8.5
$N_1 = 10$	16.5	$N_2 = 14$	123.5

The two samples in the table above are unequal, which may happen when using an independent groups design.

▼ Tables of critical values of *U* ($p \leq 0.05$)

CRITICAL VALUES FOR A ONE-TAILED TEST

N_2 \ N_1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2				0	0	0	1	1	1	1	2	2	2	3
3		0	0	1	2	2	3	3	4	5	5	6	7	7
4		0	1	2	3	4	5	6	7	8	9	10	11	12
5	0	1	2	4	5	6	8	9	11	12	13	15	16	18
6	0	2	3	5	7	8	10	12	14	16	17	19	21	23
7	0	2	4	6	8	11	13	15	17	19	21	24	26	28
8	1	3	5	8	10	13	15	18	20	23	26	28	31	33
9	1	3	6	9	12	15	18	21	24	27	30	33	36	39
10	1	4	7	11	14	17	20	24	27	31	34	37	41	44
11	1	5	8	12	16	19	23	27	31	34	38	42	46	50
12	2	5	9	13	17	21	26	30	34	38	42	47	51	55
13	2	6	10	15	19	24	28	33	37	42	47	51	56	61
14	2	7	11	16	21	26	31	36	41	46	51	56	61	66
15	3	7	12	18	23	28	33	39	44	50	55	61	66	72

CRITICAL VALUES FOR A TWO-TAILED TEST

N_2 \ N_1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2							0	0	0	0	1	1	1	1
3				0	1	1	2	2	3	3	4	4	5	5
4			0	1	2	3	4	4	5	6	7	8	9	10
5		0	1	2	3	5	6	7	8	9	11	12	13	14
6		1	2	3	5	6	8	10	11	13	14	16	17	19
7		1	3	5	6	8	10	12	14	16	18	20	22	24
8	0	2	4	6	8	10	13	15	17	19	22	24	26	29
9	0	2	4	7	10	12	15	17	20	23	26	28	31	34
10	0	3	5	8	11	14	17	20	23	26	29	33	36	39
11	0	3	6	9	13	16	19	23	26	30	33	37	40	44
12	1	4	7	11	14	18	22	26	29	33	37	41	45	49
13	1	4	8	12	16	20	24	28	33	37	41	45	50	54
14	1	5	9	13	17	22	26	31	36	40	45	50	55	59
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64

For any N_1 and N_2 observed value of *U* must be EQUAL TO or LESS THAN the critical value in this table for significance to be shown.

Source: R. Runyon and A. Haber (1976). *Fundamentals of behavioural statistics* (third edition). Reading, Mass: McGraw-Hill.

INFERENTIAL TESTS: WILCOXON *T* TEST

▲ Frank Wilcoxon (1892–1965)

The final inferential test you need to study is one that is appropriate for tests of difference where pairs of data are related, such as when a **repeated measures design** has been used. Each participant is tested twice and their scores are compared to see if there is any difference. **Matched pairs** is also a related design – where there are two groups of participants, but each participant in one group is matched with a participant in the other group on key variables, so in a sense it is like testing the same person twice.

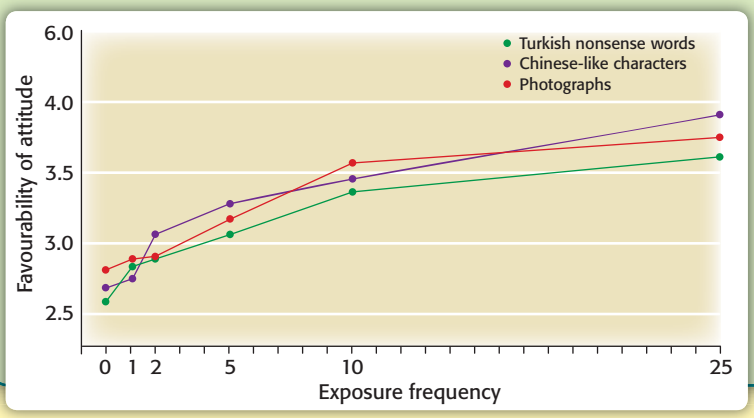
◀ Which person is more likeable? Zajonc (1968) showed that our liking for faces and objects was a function of how often we saw the face/object. Of course it might be that one of the faces is actually more likeable – you can control this extraneous variable by varying which photo you use as the 'most frequent' condition with different participants.

THE MERE EXPOSURE EFFECT

There is a saying that 'familiarity breeds contempt', but psychological research has found that the opposite is generally true – we come to like things because of their familiarity. For example. people generally like a song more after they have heard it a few times, and advertisements often aim to increase our liking for a product through repeated exposure. Things that are familiar are less strange and threatening and thus more likeable.

Zajonc (pronounced 'zie-unts') conducted various experiments to demonstrate the *mere exposure effect*. For example, in one study Zajonc (1968) told participants that he was conducting a study on visual memory and showed them a set of photographs of 12 different men (face only). Each photograph was shown for two seconds only. At the end, participants were asked to rate how much they liked the 12 different men on a scale from 0 to 6. The key element of the study is that some photos were shown more often than others, for example one photo appeared 25 times whereas another only appeared once.

Overall, the frequencies were 0, 1, 2, 5, 10 and 25. The same experiment was repeated with invented Chinese symbols and also with Turkish words. The results are shown in the graph on the left.



DO-IT-YOURSELF

You can replicate this study but don't need to have all six conditions. The final analysis can involve just comparing two of the stimuli – one frequent and one infrequent – as shown in the worked example on the facing page.

MORE DO-IT-YOURSELF IDEAS FOR STUDIES USING A REPEATED MEASURES TEST

- **Mere exposure again**. The mere exposure effect can also be used to explain the fact that people prefer pictures of themselves that are reversed as in a mirror – because that is the way you usually see yourself, and so it is more familiar (Mita *et al.*, 1977). You could take a few pictures of each participant with a digital camera and create a mirror image of each. Show them the photographs and record the ratings (on a scale of 1 to 5) for each photograph.
- **Right brain/left brain**. If you perform two tasks that involve the same brain hemisphere you should be slower on both tasks than if performing two tasks that involve the right and left hemispheres separately. For example tap your right finger while reading a page from a book (both involve the left hemisphere). Then repeat the finger-tapping without any reading. On each occasion count how many finger taps you manage in 30 seconds and compare these scores.
- **Smiling makes you happy**. You might think that you smile because you are feeling happy, but psychological research shows it works the other way round too, i.e. you become happy because you are smiling. Laird (1974) told participants to contract certain facial muscles so he could measure facial muscular activity using electrodes. Participants who were made to smile rated cartoons as funnier than those who were made to produce a frown. You could replicate this by, for example, asking people to rate ten cartoons for humour. Ask them to smile for the first five, and frown for the next five.

WHEN TO USE THE WILCOXON T TEST

- The hypothesis predicts a **difference** between two sets of data.
- The two sets of data are pairs of scores from one person (or a matched pair) = **related**.
- The data are **ordinal** or **interval** (i.e. not nominal). See page 287 for an explanation.

THE WILCOXON T TEST – A WORKED EXAMPLE

STEP 1. State the alternative and null hypothesis

Alternative hypothesis: Participants rate the more frequently seen face as more likeable than the less frequently seen face. (This is a directional hypothesis and therefore requires a one-tailed test.)

Null hypothesis: There is no difference in the likeability score for faces seen more or less often.

STEP 2. Record the data, calculate the difference between scores and rank

- Once you have worked out the difference, rank from low to high, ignoring the signs (i.e. the lowest number receives the rank of 1).
- If there are two or more of the same number (tied ranks), calculate the rank by working out the mean of the ranks that would have been given.
- If the difference is zero, omit this from the ranking and reduce N accordingly.

Participant	Likeability for *more* frequently seen face	Likeability for *less* frequently seen face	difference	rank
1	5	2	3	9.5
2	4	3	1	3
3	3	3	omit	
4	6	4	2	6.5
5	2	3	−1	3
6	4	5	−1	3
7	5	2	3	9.5
8	3	4	−1	3
8	6	3	3	9.5
10	4	6	−2	6.5
11	5	2	3	9.5
12	3	4	−1	3

STEP 3. Find observed value of T

T = the sum of the ranks of the less frequent sign.

In this case the less frequent sign is minus, so T = 3 + 3 + 3 + 6.5 + 3 = 18.5

STEP 4. Find critical value of T

N = 11 (one score omitted). The hypothesis is directional, therefore a one-tailed test is used.

Look up critical value in table of critical values (see above right).

For a one-tailed test, N = 11, the critical value of T ($p \leq 0.05$) = 13

STEP 5. State the conclusion

As the observed value (18.5) is greater than the critical value (13) we must accept the null hypothesis (at $p \leq 0.05$) and conclude that there is no difference in the likeability score for faces seen more or less often.

▼ Table of critical values of *T* ($p \leq 0.05$)

N =	One-tailed test	Two-tailed test
5	T≤0	
6	2	0
7	3	2
8	5	3
9	8	5
10	11	8
11	13	10
12	17	13
13	21	17
14	25	21
15	30	25
16	35	29
17	41	34
18	47	40
19	53	46
20	60	52
21	67	58
22	75	65
23	83	73
24	91	81
25	100	89
26	110	98
27	119	107
28	130	116
29	141	125
30	151	137
31	163	147
32	175	159
33	187	170

Observed value of *T* must be EQUAL TO or LESS THAN the critical value in this table for significance to be shown.

Source: R. Meddis (1975). *Statistical handbook for non-statisticians*. London: McGraw Hill.

CAN YOU...? No.16.10

...1 Identify the maximum observed value of T that would be required for significance with a two-tailed test with 25 participants.

...2 In a psychology experiment, 15 students were given a test in the morning and a similar test in the afternoon to see at what time of day they performed better. The research expected them to do better in the morning. Write an appropriate alternative and null hypothesis for this study.

...3 Invent data for this study – you need 15 pairs of scores.

...4 Explain why the Wilcoxon test would be the appropriate test to use with this data.

...5 Follow the steps outlined above to calculate T and then state the conclusion you would draw about the significance of the results.

...6 One problem with this study is that the students might do better in the afternoon because they had done a similar test in the morning. Therefore the study was conducted again using a matched pairs design. Explain how this might be done (including relevant variables that you would use for matching and explain why they are relevant).

...7 Explain how counterbalancing could be used to deal with the order effects in this study.

DESCRIPTIVE AND INFERENTIAL STATISTICS

In the examination you may be asked to identify appropriate descriptive and/or inferential statistics that could be used for a particular psychological study – a study may be described for you or you may be required to design your own study. Some examples of such questions are shown on the facing page.

On this page we present a reminder of all the different kinds of statistical methods you should be familiar with and how you can decide which statistic(s) would be appropriate in any situation. Such decisions are not always black and white, which means that you need to take relative strengths and limitations into account.

DECIDING WHICH STATISTICS TO USE

Descriptive statistics

This is a summary of the descriptive statistics covered at AS, and their strengths and limitations.

Measures of central tendency inform us about central (or middle) values for a set of data. They are 'averages' – ways of calculating a typical value for a set of data. An average can be calculated in different ways:

- The **mean** is calculated by adding up all the scores and dividing by the number of scores. It makes use of the values of all the data but can be unrepresentative of the data as a whole if there are extreme values. It is *not* appropriate for nominal data.
- The **median** is the middle *value* in an *ordered* list. It is not affected by extreme scores but is not as 'sensitive' as the mean because not all values are reflected in the median. It is *not* appropriate for nominal data.
- The **mode** is the value that is *most* common in a data set. It is the only method appropriate when the data are in categories (such as number of people who like pink) i.e. nominal data, but can be used for all kinds of data. It is not a useful way of describing data when there are several modes.

Measures of dispersion inform us about the spread of data.

- **Range** – Calculated by finding the difference between the highest and lowest score in a data set. This is easy to calculate but may be affected by extreme values.
- **Standard deviation** expresses the spread of the data around the mean. This is a more precise measure because all the values of the data are taken into account. However some characteristics of the data are not expressed, such as the influence of extreme values.

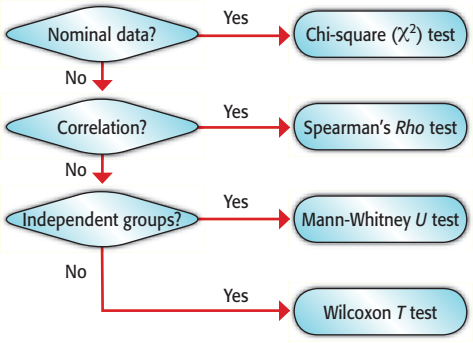
Graphs – A picture is worth a thousand words! Graphs provide a means of 'eyeballing' your data and seeing the results at a glance.

- **Bar chart** – The height of the bar represents frequency. Suitable for words and numbers i.e. all levels of measurement.
- **Scattergram** – Suitable for correlational data, a dot or cross is shown for each pair of values. If the dots form a pattern going from bottom left to top right this indicates a **positive correlation**, whereas top left to bottom right suggests a **negative correlation**. If there is no detectable pattern there is a **zero correlation**.

Inferential statistics

Inferential tests require time and patience but they are the only way to determine whether the results of a study are significant i.e. a real effect has been demonstrated as opposed to a chance pattern that *looks* meaningful.

There are a large number of inferential tests which are used by researchers and statisticians; however you only need be concerned with the four in the specification. When deciding which test is appropriate in any situation, you can ask yourself the three questions in the diagram below:



Justifying your choice – In an examination question, you may be asked to justify the choice of inferential test, either one that has been conducted on some data or one you have chosen. Below are a variety of possible justifications that could be used. In each case, full reference to the data has been made, as well as mentioning other important criteria for deciding which test to use.

Spearman's *rho* – A test of correlation is needed, as the hypothesis predicted a correlation. The data involved ratings made by participants that are ordinal data. This means we should use Spearman's *rho* (test of correlation, ordinal data).

Chi-square – As the data have been put into categories, they are classified as nominal data. The results are independent in each cell, and the *expected* frequencies in each cell are greater than 4. The appropriate inferential test to use is therefore a chi-square test (test of association, independent groups, nominal data).

Mann-Whitney – A test of difference is required because the hypothesis predicts that there will be a difference between the two groups. The design is independent groups, as participants were allocated to one of two treatment groups, and the data were scores on a test (ordinal data). Therefore the Mann-Whitney test is suitable (test of difference, independent groups, ordinal data).

Wilcoxon – A test of difference is required because the hypothesis predicts that there will be a difference between the two conditions. The design is repeated measures as all the participants were tested twice. The data were scores on a memory test, which are interval data. Therefore a Wilcoxon test was chosen (test of difference, related groups, interval data).

...1 In the following research studies suggest a suitable alternative hypothesis for the study; briefly explain how you might conduct the study; invent a hypothetical set of data that might be produced and finally select an appropriate inferential test, justifying the reason for your choice.

(a) An experiment where reaction times are compared for each participant before and after drinking coffee.
(b) A study looking at whether old or young people watch more violence on TV.
(c) An investigation to see if reaction time is related to age.
(d) An experiment to compare stress levels in doctors and nurses.
(e) A study where two groups of participants were matched on memory ability. Each group used a different revision technique to learn a topic and then their performances were compared.
(f) A study to see whether people who have a pet are happier than those who don't.

...2 For each of the following data sets identify an appropriate measure of central tendency and dispersion and justify your choice.

(a) 8, 11, 12, 12, 14, 15, 16, 16, 17, 19, 22, 27
(b) 15, 17, 21, 25, 28, 29, 32, 34, 25, 35, 38, 41, 45
(c) yes, yes, no, no, no, yes, no, no

...3 A student designed an experiment that used a repeated measures design to investigate obedience to male and female teachers. The student decided to do this by observing how pupils behaved with different teachers. She asked various friends to record student behaviours in their classrooms.

(a) State a possible directional hypothesis for this study.
(b) Suggest **two** possible extraneous variables that might be a problem for this study and describe the possible effects they could have.
(c) Suggest **three** behavioural categories that might be used to record the students' behaviours.
(d) Identify the sampling method that is likely to have been used in this study and explain why it would be chosen.
(e) Suggest some appropriate statistical measures that could be used when analysing the data (descriptive and inferential). Justify your choice.
(f) The student was asked to write a report about her study. Outline the sections that it is conventional to have in a psychological report, and give a brief description of what should be included in each section (see page 278).

...4 A psychologist designs a set of questions to collect data about smokers' and non-smokers' attitudes to smoking.

(a) Write **one** open and **one** closed question he might use. [2]
(b) For each question the psychologist would like to summarise the answers that are given. Suggest **two** ways that data could be summarised from the questions you have written.
(c) Suggest **one** advantage and **one** disadvantage of presenting the questions in writing rather than conducting face-to-face interviews.
(d) Why would standardised instructions be necessary?
(e) What inferential test might be used in this study? Justify your choice.
(f) How might demand characteristics be a problem in this study?

...5 A local hospital decides to have mixed wards rather than separate wards for men and women. Before introducing this new scheme to all wards, the hospital management decides to compare the effects of mixed versus separate wards on patient wellbeing. The hospital employs a psychologist to conduct a study on patients in mixed versus single-sex wards in terms of happiness and health. Health outcomes could be determined by looking at whether patients recover more quickly in one type of ward than another, and also at whether they have better signs of health (e.g. lower blood pressure).

(a) Identify the independent variable in this study.
(b) Identify **two** possible dependent variables and suggest how you could operationalise them.
(c) Write an appropriate alternative hypothesis for this study.
(d) (i) Identify the experimental design used in this study.
(ii) Describe **one** disadvantage of this design in the context of this study.
(iii) Explain **one** way of dealing with this disadvantage.
(e) (i) The psychologist uses a Mann-Whitney *U* test to check whether there is a significant difference between the recovery rates of 12 patients on mixed and 12 patients on single-sex wards.
(ii) Explain why this test was chosen.
(iii) The test produced an observed (calculated) value of U=29. Using the table below, explain whether the results support the hypothesis that you proposed in part (b).

▼ Critical values of *U* at 5% level ($p \leq 0.05$) for a two-tailed test

		N_1					
		10	11	12	13	14	15
N_2	10	23	26	29	33	36	39
	11	26	30	33	37	40	44
	12	29	33	37	41	45	49
	13	33	37	41	45	50	54
	14	36	40	45	50	55	59
	15	39	44	49	54	59	64

For any N_1 and N_2, the observed value of *U* must be EQUAL TO or LESS THAN the critical value in this table for significance to be shown.

(f) Blood pressure readings are recorded for patients on mixed and single-sex wards. Suggest three appropriate descriptive statistics that could be used to represent the data. Justify your choice.
(g) An alternative way to find out about people's preferences for hospital accommodation would be to conduct a survey. Write a plan for conducting this study. You should include sufficient details to permit replication, for example a hypothesis, details of design and procedure, sampling and ethical issues.

Statistical joke: 'I read that there is about one chance in one million that someone will board an airplane carrying a bomb, and so I started carrying a bomb with me on every flight I take. The way I figure it, the odds against two people having a bomb on the same plane are one in a trillion.'

ANALYSIS AND INTERPRETATION OF QUALITATIVE DATA

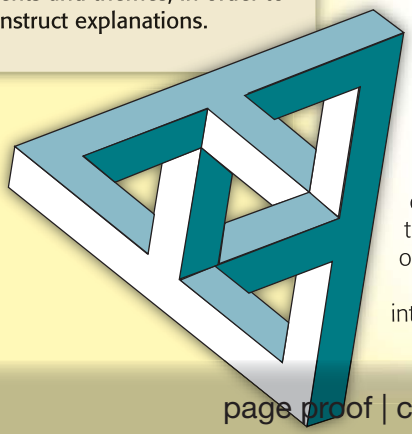
Descriptive and inferential statistics are methods of **quantitative data** analysis i.e. methods of analysing numerical data. You are also familiar with **qualitative data**, where information is gathered in a non-numeric form. Both quantitative and qualitative data may concern thoughts and feelings or any aspect of behaviour; the difference lies in the form the data takes. Quantitative and qualitative data may be produced in interviews, observational studies and case studies. Quantitative analysis involves counting responses or occurrences whereas qualitative analysis is concerned with interpreting the *meaning* of data i.e. quality rather than quantity.

Quantitative data *is data that represents how much or how long, or how many, etc. there are of something, i.e. behaviour that is measured in numbers or quantities.*

Qualitative data *is essentially anything that is not in numerical form, for example what people say or write. Qualitative data can't be counted but it can be summarised.*

SOME KEY POINTS ABOUT QUALITATIVE RESEARCH AND ANALYSIS

- Qualitative researchers believe that the traditional quantitative methods used by psychologists do not produce results that are applicable to everyday life.
- Qualitative methods emphasise subjectiveness because they aim to represent the world as seen by the individual.
- In order to produce subjective information the qualitative researcher asks broad questions that allow a respondent to answer in their own words, or observes behaviour directly or indirectly (e.g. through things that people have written or drawn).
- The data sets produced in qualitative research tend to be very large, though the samples may be quite small compared with those used in quantitative approaches.
- Qualitative data usually cannot be reduced to numbers. If a researcher is trying to produce numbers then he or she is probably not engaged in qualitative data analysis.
- The data can be examined for differences and similarities across different cases, times, events and themes, in order to construct explanations.



METHODS OF QUALITATIVE DATA ANALYSIS

There are a wide variety of approaches or methods used by qualitative researchers, such as *discourse analysis* (studying written or oral discourses), *ethnography* (observing people in their natural environments), and *Interpretative Phenomenological Analysis* (IPA – understanding how people make sense of their own experiences).

Coding

Coding is the process of identifying categories, themes, phrases or keywords that may be found in any set of data. For example when analysing observational data, the researcher identifies a number of categories and then allocates each individual observation to one of the categories. Similarly, when analysing the transcript of an interview, the researcher identifies a variety of themes (such as feeling upset or thinking about the future) and then works through the entire text annotating each sentence of the interview.

Coding is not a superficial categorisation, but a thoughtful process aimed at trying to understand the meaning of the data. The categories or themes are decided upon in one of two ways:

1. Top-down approach (**thematic analysis**) – codes represent ideas and concepts from an existing theory/explanation. For example the clinical characteristics of schizophrenia may be used as categories to code self-descriptions from patients diagnosed with schizophrenia.
2. Bottom-up approach (**grounded theory**) where the codes/categories emerge from the data. Thus codes remain grounded in the observations rather than being generated beforehand by existing views. This is popular in an area that has not been well researched or in order to develop new insights.

The analysis of qualitative data requires repeated reviewing of the data in order to consider further categories, to re-assess meaning and to re-assign codes.

Summarising the data

Later the behavioural categories can be used to summarise the data. For example, the categories or themes may be listed, or examples of behaviour within the category may be represented using quotes from participants, or descriptions of typical behaviours in that category. It is also possible to count frequency of occurrences in each category, thus qualitative data is turned into quantitative data. Finally a researcher may draw conclusions.

Validity and reflexivity

The traditional approach in psychology claims that there is one real world, and quantitative research seeks to discover that reality or 'truth' – validity is a measure of the extent to which that has been achieved. The qualitative approach denies the existence of any one 'real' world, i.e. each individual's subjective perspective is 'reality'. Qualitative researchers acknowledge the need for **reflexivity** – the recognition that a researcher's attitudes, biases etc. have an unavoidable influence on the research they are conducting. The impact of reflexivity cannot be avoided but it can be monitored and reported.

The **validity** of qualitative research findings may be demonstrated using **triangulation**, comparing the results from a variety of different studies of the same thing or person. The studies are likely to have used different methodologies. If the results agree, this supports their validity. If the results differ, this can lead to further research to enhance our understanding.

Reliability is a component of validity and can be checked, for example by looking at inter-rater reliability when more than one person has coded the data.

EXAMPLES OF QUALITATIVE ANALYSIS AND INTERPRETATION

A qualitative analysis

A Finnish study considered the role of the family in adolescents' experiences with friends. Joronen and Åstedt-Kurki (2005) conducted semi-structured interviews with 19 adolescents aged 12–16, using questions such as: 'What does your family know about your peers?' and 'How is your family involved in your school activities?' These interviews produced 234 pages of notes which were analysed using a qualitative content analysis.

1 All answers to the same questions were placed together.

2 Each statement was compressed into a briefer statement and given an identifier code.

3 These statements were compared with each other and categorised so that statements with similar content were placed together and a category identified (a thematic analysis).

4 The categories were grouped into larger units producing eight main categories, for example
- *Enablement* e.g. 'Yeah, ever since my childhood we've always had lots of kids over visiting.' (Girl, 15 years)
- *Support* e.g. 'They [family members] help if I have a test by asking questions.' (Boy, 13 years)
- *Negligence* e.g. 'My sister is not at all interested in my friends.' (Girl, 16 years)

One of the conclusions drawn from this study is that schools should pay more attention to the multiple relationships that determine an adolescent's behaviour.

Collaborative research

Collier *et al.* (2005) conducted a study of how people form relationships by arranging for 10 female undergraduates, previously unacquainted, to be randomly assigned to partners and record their thoughts and feelings when forming a relationship. An essential element of this study was that the participants collaborated fully in the research; their ability to understand and reflect on their own experience means that they were the 'experts' with the researcher acting more as a 'facilitator'. Such **collaborative research** is typical of the aims of qualitative approaches.

The meetings between partners were recorded on audiotape. Each woman was interviewed over the course of the study about her thoughts, and each woman kept a weekly diary. Three of the relationships were selected for analysis because they represented rather different experiences of intimacy. The analysis looked at various aspects of relationships, for example:

- *Similarities and differences* i.e. the extent to which partners were similar to or different from each other and how these perceptions changed over time and affected relationship formation.
- *Self-disclosure* (telling someone personal information). Past research suggests that disclosure from one partner should lead to disclosure from another, but this study found that this only occurred if the disclosure communicated trust.

COMPARING QUANTITIATIVE AND QUALITATIVE DATA

	Advantages	*Weaknesses*
Quantitative data	• Easier to analyse because data in numbers. • Produces neat conclusions.	• Oversimplifies reality and human experience (statistically significant but humanly insignificant).
Qualitative data	• Represents the true complexities of human behaviour. • Gains access to thoughts and feelings which may not be assessed using quantitative methods with closed questions. • Provides rich detail.	• More difficult to detect patterns and draw conclusions. • Subjective analysis can be affected by personal expectations and beliefs (though quantitative methods may only appear to be objective but are equally affected by bias).

DO-IT-YOURSELF IDEAS FOR STUDIES USING QUALITATIVE DATA ANALYSIS

- **Analysis of a questionnaire**. Design a questionnaire on any topic of your own choosing with a number of open-ended questions. Analyse the questions using a top-down or bottom-up approach. Summarise the findings using quotes from respondents and present some conclusions.
- **Advertisements** on TV or in magazines/newspapers. You might consider how men and women are represented in advertisements. Manstead and McCulloch (1981) looked at ads on British TV (170 ads over a one-week period, ignoring those that contained only children and animals). In each ad they looked at what the central adult figure was doing, and recorded frequencies in categories such as whether men or women were cast in a dependent role, presented the central argument, were shown at home or at work, etc. You could produce your own categories, based on observations, and present a qualitative analysis of the ads you look at (as opposed to a frequency analysis that would be quantitative).
- **Mental disorders** – There are various websites that publish self-descriptions of people with mental disorders, such as eating disorders or depression. You could investigate by either taking a thematic approach (look for symptoms that are typical of the disorder) or a grounded theory approach (develop your own clinical characteristics by reading individual reports).

CAN YOU...? No.16.12

...**1** Describe **one** advantage and one disadvantage of qualitative analysis over quantitative analysis.

...**2** Select **one or more** studies you are familiar with and give examples of both quantitative and qualitative data collected in this study.

...**3** Explain how the opinions of the participants might be represented in a qualitative analysis.

...**4** Explain the concept of triangulation and how it is useful.

...**5** Explain why qualitative researchers are less concerned about conducting 'valid' research.

...**6** Explain the difference between thematic analysis and grounded theory.

...**7** On the left is a qualitative content analysis. In what way is it a content analysis? In what way is it qualitative?

...**8** The second study (by Collier *et al.*) is an example of collaborative research. Outline **one** conclusion that could be drawn from this research.

...**9** A researcher conducts a case study of a child who has spent several long periods of time in hospital for a bone disorder. Suggest how this study might be conducted including some themes which might be examined and how the data might be summarised at the end of the study.

CHAPTER SUMMARY

This section builds on the knowledge and skills developed at AS level. There are a number of concepts that were included in our AS research methods chapter, that are not covered in this chapter. The concepts are listed below and you can find their definitions in the glossary index at the back of this book:

- **attrition**
- **cohort effects**
- **confederate**
- **control condition/group**
- **covert observations**
- **cross-sectional study**
- **difference studies**
- **double blind**
- **effect size**
- **experimental condition/group**
- **experimental design**
- **experimental realism**
- **Hawthorne effect**
- **imposed etic**
- **intervening variable**
- **interviewer bias**
- **participant effects**
- **participant variables**
- **presumptive consent**
- **protection from harm**
- **quasi-experiments**
- **random allocation**
- **role play**
- **single blind**
- **situational variables**
- **structured (systematic) observations**
- **unstructured interview**
- **unstructured observations**

SCIENTIFIC METHOD

SCIENCE

MAJOR FEATURES OF SCIENCE

- Empiricism
- Objectivity
- Replicability
- Control
- Theory construction

SCIENTIFIC PROCESS

- Induction – reasoning from particular to general.
- Deduction – reasoning from general to particular.

COMMENTARY – IS PSYCHOLOGY SCIENTIFIC?

- Scientific research is desirable.
- Psychology shares the goals of science.
- Kuhn – no single paradigm.
- Lack of objectivity and control leads to experimenter bias and demand characteristics.

COMMENTARY – ARE GOALS OF SCIENCE APPROPRIATE?

- Nomothetic versus idiographic.
- Qualitative research – triangulation.

SYNOPTIC LINKS

Scientific approach is:

- reductionist – reduces complex phenomena to simple ones; and
- determinist – searches for causal relationships.

VALIDATING KNOWLEDGE

PEER REVIEW

Serves three main purposes:

- Allocation of research funding
- Publication in scientific journals
- Research Assessment Exercise.

Research published on the Internet requires new solutions.

COMMENTARY

- May be an unachievable ideal.
- Anonymity allows honesty and objectivity.
- Publication bias favours positive results.
- May lead to preservation of the status quo.

CONVENTIONS OF SCIENTIFIC REPORTING

- Abstract – summary of study.
- Introduction/aim – literature review and research intentions.
- Method – procedures and design of study.
- Results – descriptive and inferential statistics.
- Discussion – outcomes and implications of study.
- References.

SYNOPTIC LINKS

- Some changes in science are not logical changes but represent a shift in perspective (paradigm shift).
- Burt research – an example of scientific fraud.

DATA ANALYSIS

PROBABILITY AND SIGNIFICANCE

PROBABILITY AND SIGNIFICANCE

- Probability = likelihood that a pattern of results could arise by chance.
- If probability extremely unlikely, then result is statistically significant.
- Inferential tests determine whether chance or real trend in data.
- Probability levels represent acceptable level of risk (e.g. $p \leq 0.05$) of making a Type 1 error.
- More important research, more stringent significance levels.
- Type 1 error = null hypothesis rejected when true.
- Type 2 error = null hypothesis accepted when false.

INFERENTIAL TESTS

- Different research designs require different tests.
- Different tests for different levels of measurement (nominal, ordinal, interval, ratio).
- Tests yield *observed* values, and then compared to *critical* values to determine significance level.
- One-tailed test = directional hypothesis.
- Two-tailed test = non-directional hypothesis.

INFERENTIAL TESTS

SPEARMAN'S *RHO*

Used when:

- Hypothesis predicts correlation between two variables.
- Each person is measured on both variables.
- Data is at least ordinal (i.e. not nominal).

MANN-WHITNEY *U*

Used when:

- Hypothesis predicts difference between two sets of data.
- Independent groups design.
- Data at least ordinal (i.e. not nominal).

CHI-SQUARE

Used when:

- Hypothesis predicts differences between two conditions or association between two variables.
- Data is independent.
- Data in frequencies (nominal).
- Expected frequencies in each cell must not fall below 5.

WILCOXON *T*

Used when:

- Hypothesis predicts difference between two sets of data.
- Related design (repeated measures or matched pairs).
- Data at least ordinal (i.e. not nominal).

DESCRIPTIVE STATISTICS

CENTRAL TENDENCY

- Indicates typical or 'average' score.
- Mean = sum of all scores divided by number of scores. Unrepresentative if extreme scores.
- Median = middle value in ordered list of scores. Not affected by extreme scores but not as sensitive as mean.
- Mode = most common value. Not useful if there are many modes in a set of scores.

DESIGNING INVESTIGATIONS

RESEARCH METHODS

EXPERIMENTS
- IV varied to see effect on DV.
- Laboratory experiment – high on internal validity, low on external validity.
- Field experiment – more natural environment but more issues of control than laboratory experiment.
- Natural experiment – uses naturally occurring IVs but cannot conclude causality.
- Experimental designs – repeated measures, independent groups, matched pairs.

SELF-REPORT METHODS
- Questionnaires and interviews.
- Structured interviews – more easily repeated.
- Unstructured interviews – questions that evolve are dependent on answers given.
- May involve open (respondent provides own answer) or closed (respondent chooses specific answer) questions.
- Main problem: social desirability bias.

OBSERVATIONAL STUDIES
- Observing behaviour through behavioural categories.
- Sampling methods – time and event sampling.
- Open to subjective bias – observations affected by expectations.

CORRELATIONAL ANALYSIS
- Concerned with relationship between two variables.
- Does not demonstrate causality.
- Other variables may influence any measured relationship.

CASE STUDIES
- Detailed study of individual, institution or event.
- Generally longitudinal, following individual or group over time.
- Allows study of complex interaction of many variables.
- Difficult to generalise from specific cases.

DESIGN ISSUES

RELIABILITY
- Experimental research – allows for replication of study.
- Observations – inter-observer reliability can be improved through training.
- Self-report – internal reliability (split-half) and external reliability (test-retest).

VALIDITY
- Internal validity – does study test what it was intended to test?
- External validity – can results be generalised to other situations and people?
- Laboratory experiments not necessarily low in external validity.
- If low in mundane realism, reduces generalisability of findings.
- In observations, internal validity affected by observer bias.
- Self-report techniques, issues of face and concurrent validity.

SAMPLING TECHNIQUES
- Opportunity – most easily available participants.
- Volunteer – e.g. through advert, but subject to bias.
- Random – all members of target population must have equal chance of selection.
- Stratified and quota – different subgroups within sample, leads to more representative sample.
- Snowball – researcher directed to other similar potential participants.

ETHICS

ETHICAL ISSUES WITH HUMANS
- Informed consent and deception.
- Harm – what constitutes too much?

CODE OF CONDUCT
- Respect for worth and dignity of participants.
- Right to privacy, confidentiality, informed consent and right to withdraw.
- Intentional deception only acceptable in some circumstances.
- Competence – retaining high standards.
- Protection from harm and debriefing.
- Integrity – being honest and accurate in reporting.
- Use of ethical guidelines in conjunction with ethical committees.
- Socially sensitive research – potential social consequences for participants.

ETHICAL ISSUES WITH NON-HUMANS
- Reasons for animals use – offers opportunity for greater control and objectivity; can't use humans; physiological similarities.
- Moral issues – sentience (experience pain and emotions).
- Specieism – form of discrimination against non-human species.
- Animal rights – Regan (1984), no animal research is acceptable.
- Do animals have rights if they have no responsibilities?
- Animal research subject to strict legislation (Animals Act; BPS guidelines).
- The 3Rs – Reduction, Replacement, Refinement.

DESCRIPTIVE STATISTICS cont'd

MEASURES OF DISPERSION
- Indicate spread of scores.
- Range = difference between highest and lowest score. Not representative if extreme scores.
- Standard deviation = spread of data around mean. Precise measure but influence of extreme scores not taken into account.

GRAPHS
- Bar chart = illustration of frequency, height of bar represents frequency.
- Scattergram = illustration of correlation, suitable for correlational data. Indicates strength of correlation and direction (positive or negative).

QUALITATIVE DATA

KEY POINTS
- *Quantitative* methods not relevant to 'real life'.
- *Qualitative* represents world as seen by individual.
- Emphasises collection of subjective information from participant.
- Data sets tend to be large.
- Qualitative data cannot be reduced to numbers.
- Can be examined for themes.

METHODS OF ANALYSIS
- Coding using top-down approach (thematic analysis) = codes represent ideas/themes from existing theory.
- Coding using bottom-up approach (grounded theory) = codes emerge from data.
- Behavioural categories used to summarise data.
- Reflexivity indicates attitudes and biases of researcher.
- Validity demonstrated by triangulation.
- Reliability checked by inter-rater reliability.

QUANTITATIVE VERSUS QUALITATIVE
Quantitative:
- Easy to analyse and produces neat conclusions

But:
- Oversimplifies reality and human experience.

Qualitative:
- Represents true complexities of behaviour through rich detail of thoughts, feelings etc.

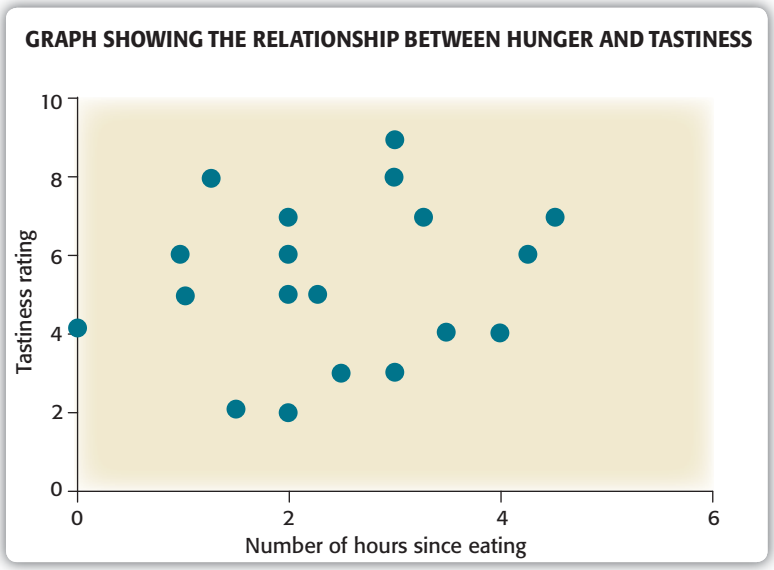
But:
- More difficult to detect patterns and subject to bias of subjectivity.

EXAM QUESTION WITH STUDENT ANSWER

Question

There is a saying that 'hunger is the best cook'. A psychologist decided to test the relationship between hunger and the tastiness of food. He prepared a dish of scrambled eggs and toast for each participant. Before they started to eat he asked them how long it was since they had last eaten. After they had eaten the meal he asked them to rate the tastiness of the meal on a scale of 1 to 10 where 10 is very tasty.

He plotted his findings as shown in the graph on the right. The correlation coefficient is 0.15.



(a) How were hunger and tastiness operationalised? *(2 marks)*

(b) This study uses a correlational analysis. Describe one weakness of using this method of analysis in this study. *(2 marks)*

(c) (i) Describe one possible threat to the validity of this study. *(2 marks)*

(ii) Explain how the psychologist could deal with this problem. *(2 marks)*

(d) Tastiness is measured using a rating scale. How could you check the reliability of this scale? *(2 marks)*

(e) (i) The scores for hunger were 2, 2, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8, 8, 9 (units of time without food). Suggest one suitable method of central tendency to use with this data. *(1 mark)*

(ii) Explain how you would calculate this measure of central tendency. *(2 marks)*

(iii) Describe one advantage of this method of central tendency. *(2 marks)*

(f) Name an appropriate statistical test for analysing this data. *(1 mark)*

(g) With reference to the graph above, describe the relationship between tastiness and hunger. *(2 marks)*

(h) (i) Write a suitable null hypothesis for this study. *(2 marks)*

(ii) The psychologist tested a non-directional hypothesis and there were 18 participants. Identify the observed value and the critical value for this data, using the table of critical values shown below. *(2 marks)*

(iii) State whether the null hypothesis should be rejected or accepted. *(1 mark)*

▼ Table of critical values of *rho* at 5% level ($p \leq 0.05$)

N =	One-tailed test	Two-tailed test
13	0.484	0.560
14	0.464	0.538
15	0.443	0.521
16	0.429	0.503
17	0.414	0.485
18	0.401	0.472
19	0.391	0.460

Observed value of *rho* must be EQUAL TO or GREATER THAN the critical value in this table for significance to be shown.

(i) The psychologist wished to have his work published in an academic journal. Outline the sections that would be likely to be present in the report that he writes for the journal. *(2 marks)*

(j) The journal is peer-reviewed. Discuss the process of peer review. *(6 marks)*

(k) Subsequently, the psychologist decided to collect some qualitative data about the effects of hunger. Briefly outline how he might do this and how he might analyse the data he collects. *(4 marks)*

STUDENT ANSWER	EXAMINER COMMENTS The answers provided on the left are all worth full marks. A few notes are provided below on some of the answers.
(a) *Hunger was measured in terms of time since participants last ate a meal and tastiness was measured using a rating scale.*	**(a)** Operationalisation essentially means 'how can it be measured'.
(b) *One weakness of using a correlational analysis in this study was that you can't demonstrate that it is the hunger that actually causes the changed perception of the pictures.*	**(b)** The answer has correctly been related to this particular study rather than being a general answer about studies using correlational analysis. This is always a problem for students, who will lose marks for answers that are not contextualised.
(c) **(i)** *One threat to validity is that some of the participants might not think scrambled eggs were very tasty no matter how hungry they were.* **(ii)** *You could deal with this by showing people about 10 pictures of different meals so there would be some they liked. This would give a better measurement of perception of food.*	**(c)** There is no single right answer to a question like this – whatever answer is chosen, it is important to give a clear explanation to the examiner so that the answer makes sense.
(d) *You could assess reliability using the test-retest method where people are asked to rate the tastiness of food and then this is repeated again about a week later with the same people who are as hungry as they were the first time. Their ratings should be the same if the measure is reliable.*	**(d)** While it is clearly necessary to say more than 'test-retest' to gain 2 marks, it is not necessary to provide as much detail as given here (though it can be useful in research methods questions to err on the side of giving too long an answer).
(e) **(i)** *A suitable method would be the mean.* **(ii)** *You would add all the numbers up and divide by the number of scores i.e. divide in this case by 20.* **(iii)** *The advantage of this method is that it takes all the values of the numbers into account.*	**(e)** Students will not have to perform actual calculations in an exam, nor will they be asked how to calculate an inferential statistic, but could be asked to describe how to do very easy calculations. (iii) Note that the key word is 'values'. All measures of central tendency use all the numbers, but only the mean uses all the values of all the numbers.
(f) *Spearman's* rho.	**(f)** No further detail is required when asked to name a suitable test.
(g) *There is possibly a small positive correlation but it doesn't look significant.*	**(g)** As this question is worth 2 marks it is necessary to say more than just 'positive correlation' or 'zero correlation' (the correlation coefficient shows it is a slight positive correlation).
(h) **(i)** *There is no relationship between hunger (time since eating) and tastiness (rating of the scrambled eggs).* **(ii)** *The observed value is .15 and the critical value is .401.* **(iii)** *The null hypothesis should be accepted.*	**(h)** In the case of Spearman's *rho* the null hypothesis is rejected only if the observed value is greater than the critical value (this information is given under the table of critical values).
(i) *The sections would be: abstract, introduction, method (procedures), results and a discussion. There would also be references.*	**(i)** No need for more detail.
(j) *A peer review is when an expert in the field being written about reviews the article to judge its quality. This is usually unpaid and often done anonymously to encourage objectivity and honesty though, at the same time, this may have the opposite effect – some reviewers might use it as an opportunity to prevent competing researchers from publishing work. Peer reviews may be an ideal, whereas in practice there are lots of problems, for example it slows publication down and may prevent unusual, new work being published. Some people doubt whether peer review can really prevent the publication of fraudulent research. The advent of the internet means that a lot of research and academic comment is being published without official peer reviews than before, though systems are evolving on the internet where everyone really has a chance to offer their opinions and police the quality of research.*	**(j)** The number of marks available for this question (6 marks) reveal that it requires a reasonably lengthy answer (1 mark is about 25 words). Throughout the research methods questions, it is important to match the length of an answer to the marks that are available. There is no point writing a lengthy answer if a question is worth just 2 marks, but, equally, it is inadvisable to write a short paragraph when there are 6 marks available. In this question it is also important to notice to injunction 'discuss', which means describe and evaluate.
(k) *In order to collect more qualitative data the psychologist would ask people questions about what kind of things made them hungry, for example is it the way food looks or is it the smell. He would use their answers to ask further questions. He would tape-record these interviews so he could make transcripts of what people said and then he could go through these transcripts to identify key themes. It might help (improve reliability and validity) to get someone else to also analyse the data so they could compare their analyses.*	**(k)** The exam is likely to include a rather longer question about designing a study. The key is to present to the examiner as much detail as possible in the answer. There can be some benefit in scraping the barrel and including what may seem like trivial details.
[505 words]	