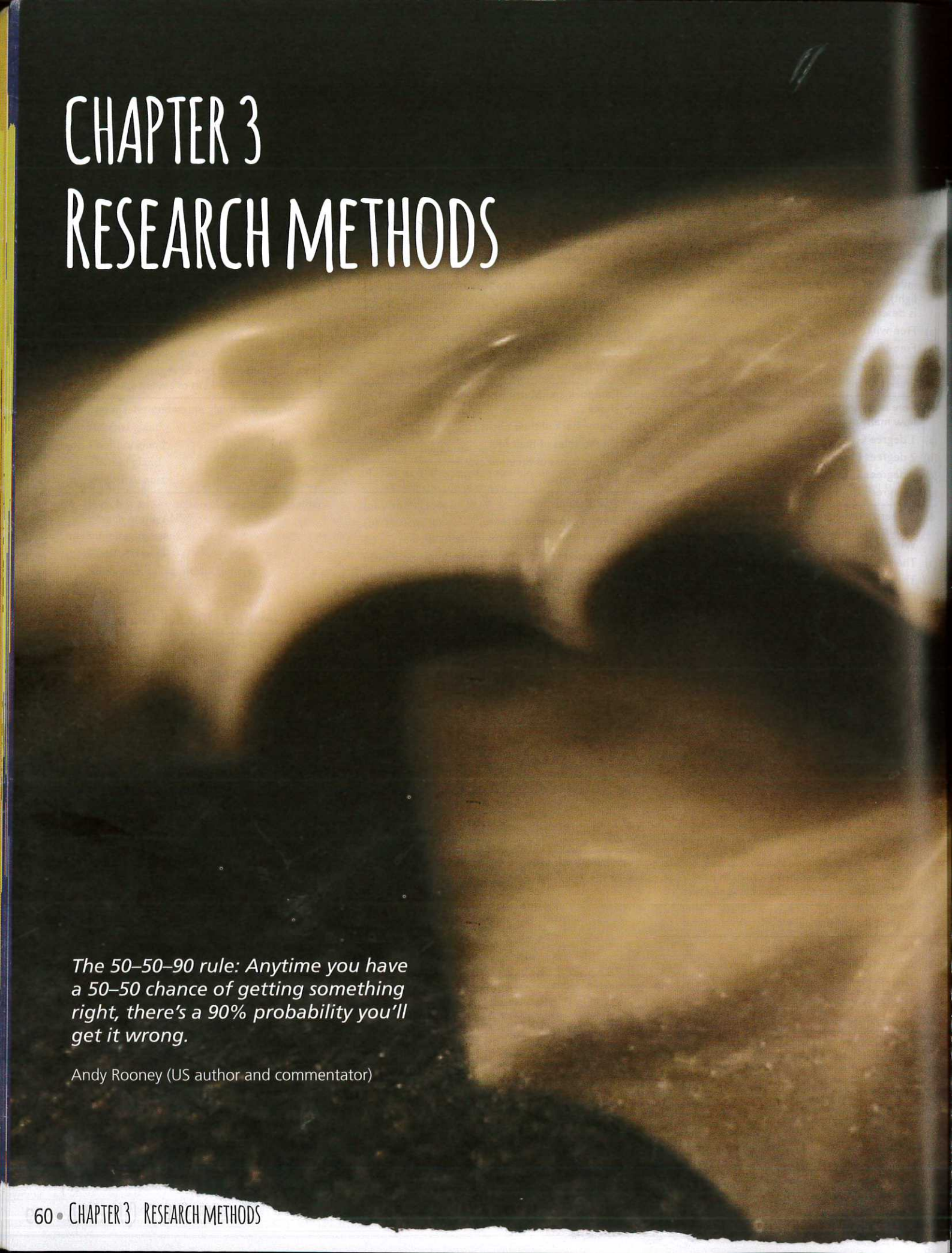
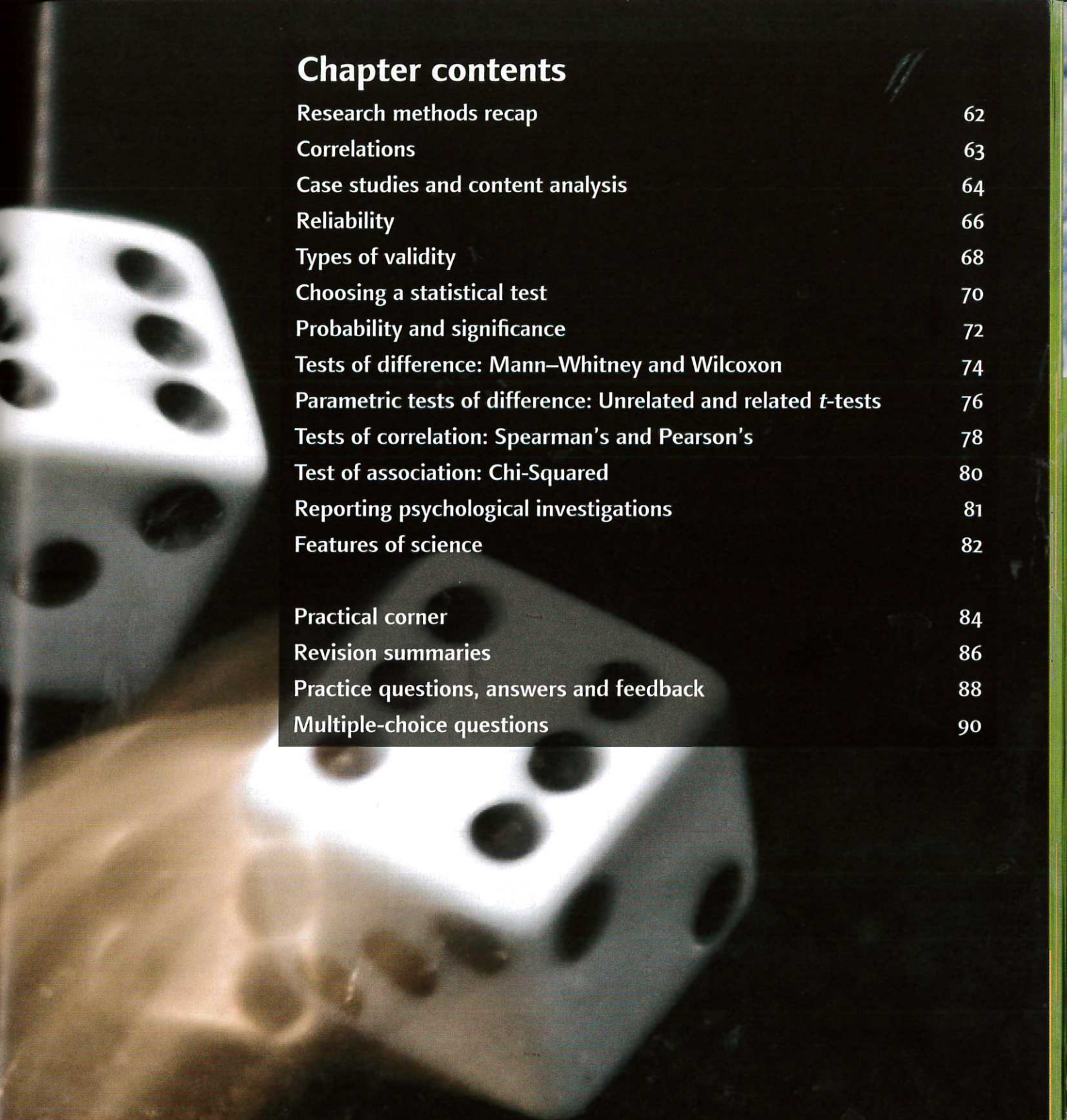
CHAPTER 3
RESEARCH METHODS



The 50–50–90 rule: Anytime you have a 50–50 chance of getting something right, there's a 90% probability you'll get it wrong.

Andy Rooney (US author and commentator)

Chapter contents



What is the probability that you will throw a six?
What is the probability it will rain next week?
What are the chances that you'll win the National Lottery next week?
What is the probability that the things psychologists discover are 'true'?
Is scientific 'proof' of something even possible?
The answers to these questions (and more) in the next few pages. Probably.

RESEARCH METHODS RECAP

THE SPECIFICATION SAYS

Students should demonstrate knowledge and understanding of the following research methods, scientific processes and techniques of data handling and analysis, be familiar with their use and be aware of their strengths and limitations.

Welcome back to Research Methods! Included on this spread is a summary of the Research Methods content that you need to know by the end of A level.

On the right is a recap of what you have already covered in Year 1. Below is a breakdown of the content for this year that is A level only.

KEY TERM

Research methods – The processes by which information or data is collected usually for the purpose of testing a hypothesis and/or a theory.

The methods bit

Overall, at least 25% of the marks in assessments for Psychology will be based on assessment of research methods. Although 50% of Paper 2 at A level will assess Research Methods, it could also be assessed in any other topic on any other paper!

Research methods – still to come ...

A level only

(You can use this to tick off topics as you complete them.)

- Case studies. Content analysis and coding. Thematic analysis. ☐
- Reliability across all methods of investigation. Ways of assessing reliability: test–retest and inter-observer; improving reliability. ☐
- Types of validity across all methods of investigation: face validity, concurrent validity, ecological validity and temporal validity; assessment of validity; improving validity. ☐
- Factors affecting the choice of statistical test, including level of measurement and experimental design. ☐
- When to use the following tests: Spearman's rho, Pearson's r, Wilcoxon, Mann–Whitney, related *t*-test, unrelated *t*-test and Chi-Squared test. ☐
- Analysis and interpretation of correlation, including correlation coefficients. ☐
- Probability and significance: use of statistical tables and critical values in interpretation of significance; Type I and Type II errors. ☐
- Reporting psychological investigations: sections of a scientific report: abstract, introduction, method, results, discussion and referencing. ☐
- Features of science: objectivity and the empirical method; replicability and falsifiability; theory construction and hypothesis testing; paradigms and paradigm shifts. ☐

Research methods – the story so far ...

AS and Year 1 Specification content

Tick off what you already know and would feel confident answering questions on in the exam. Revisit concepts if necessary.

- Aims: stating aims, the differences between aims and hypotheses. ☐
- Hypotheses: directional and non-directional. Variables and control. ☐
- Types of experiment, laboratory and field experiments; natural and quasi-experiments. Experimental designs: repeated measures, independent groups, matched pairs. ☐
- Sampling: the difference between population and sample; sampling techniques including: random, systematic, stratified, opportunity and volunteer; implications of sampling techniques, including bias and generalisation. ☐
- Ethics, including the role of the British Psychological Society's code of ethics; ethical issues in the design and conduct of psychological studies; dealing with ethical issues in research. ☐
- Observational techniques. Types of observation: naturalistic and controlled observation; covert and overt observation; participant and non-participant observation. Observational design: behavioural categories, event sampling, time sampling. ☐
- Self-report techniques. Questionnaires; interviews, structured and unstructured. Questionnaire construction, including use of open and closed questions; design of interviews. ☐
- Correlations. Analysis of the relationship between co-variables. The difference between correlations and experiments. Positive, negative and zero correlations. ☐
- Quantitative and qualitative data; the distinction between qualitative and quantitative data collection techniques. Primary and secondary data, including meta-analysis. ☐
- Descriptive statistics: measures of central tendency: mean, median, mode; calculation of mean, median and mode; measures of dispersion: range and standard deviation; calculation of range. ☐
- Mathematical content – calculation of percentages, converting a percentage to a decimal, converting a decimal to a fraction, using ratios, mathematical symbols, probability, significant figures. ☐
- Introduction to statistical testing: the sign test. ☐
- Presentation and display of quantitative data: graphs, tables, scattergrams, bar charts, histograms. Distributions: normal and skewed distributions; characteristics of normal and skewed distributions. ☐
- Pilot studies and the aims of piloting. ☐
- The role of peer review in the scientific process. ☐
- The implications of psychological research for the economy. ☐

At the end of each chapter in this book (including this one) you will find suggestions for practical investigations. You should carry out as many of these as you can to support your understanding of research methods.

CORRELATIONS

THE SPECIFICATION SAYS . . .

Analysis and interpretation of correlation, including correlation coefficients.

Correlation is not new to you – you learned about it in Year 1 of the course; the analysis and interpretation of correlation coefficients is. All correlations can be represented by a number somewhere between –1 and +1. What this number means is explained here.

KEY TERMS

Correlation – A mathematical technique in which a researcher investigates an association between two variables, called co-variables.

Correlation coefficient – A number between –1 and +1 that represents the direction and strength of a relationship between co-variables.

Apply it Methods: Interpretation of correlation coefficients

Questions

1. What sort of relationship is suggested by the following coefficients? *(5 marks)*
   (i) –.40
   (ii) +.90
   (iii) +.13
   (iv) –.76
   (v) 0
2. What are the strengths and limitations of using correlations in psychological research? *(6 marks)*

Look out for the **Apply it Methods** features in every chapter (like the one above!) so you can test your Research Methods skills . . .

Analysis and interpretation of correlations

Correlations and correlation coefficients

The term **correlation** refers to a mathematical technique which measures the relationship/association between two continuous variables (properly called **co-variables**). Such relationships are plotted on a **scattergram** where each axis represents one of the variables investigated. We shall also see, later in this chapter, how correlations/associations may be analysed using **statistical tests**.

You will study two statistical tests of correlation (see pages 78–79) each of which, when calculated, produces a numerical value somewhere between –1 and +1 known as the **correlation coefficient.** This value tells us the *strength* and *direction* of the relationship between the two variables.

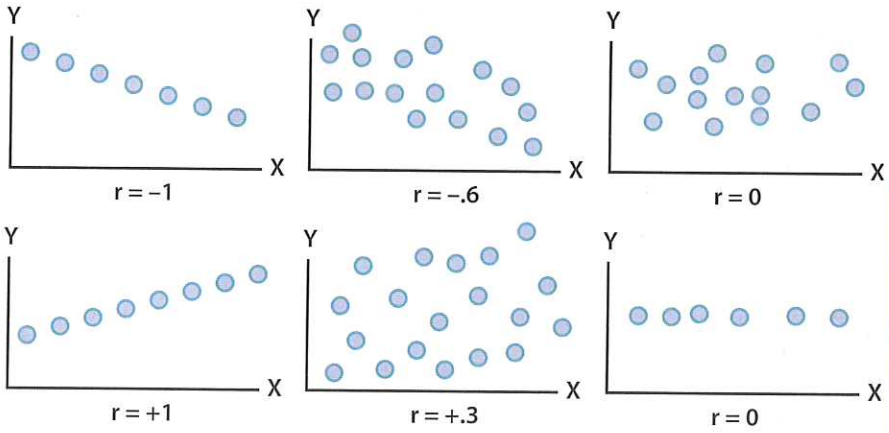
Working out what a coefficient means

As can be seen on the picture below, a value of +1 represents a *perfect* **positive correlation**, and a value of –1, a *perfect* **negative correlation**.

The closer the coefficient is to +1 or –1, the *stronger* the relationship between the co-variables is; the closer to zero, the *weaker* the relationship is.

However, it should be noted that coefficients that appear to indicate weak correlations can still be **statistically significant** – it depends on the size of the data set.

Scattergrams showing various correlation coefficients



The letter 'r' stands for correlation coefficient.

STUDY TIP

Descriptive and inferential statistics

At A level you need to be aware of the difference between **descriptive statistics** and **inferential statistics**. Descriptive statistics refers to things like graphs, tables and summary statistics (such as measures of central tendency and measures of dispersion). These are used to identify trends and analyse sets of data.

Inferential statistics refers to the use of **statistical tests** which tell psychologists whether the differences or relationships they have found are statistically significant or not. This helps decide which **hypothesis** to accept and which to reject. A **correlation coefficient** is calculated using a statistical test and, as such, is an inferential statistic.

CHECK IT

1. Explain what is meant by the term *correlation coefficient.* *[2 marks]*
2. Sketch a graph to represent a negative correlation between 'number of people in a room' and 'amount of personal space'. *[2 marks]*
3. Using an example, explain what is meant by the term *correlation.* *[2 marks]*

CASE STUDIES AND CONTENT ANALYSIS

THE SPECIFICATION SAYS...

Case studies. Content analysis and coding. Thematic analysis.

We look at two ways of investigating human behaviour not considered in Year 1: case studies and content analysis.

Case studies allow a detailed insight into a single individual, group or institution. It is a method often favoured by researchers who adopt an idiographic approach to the study of human behaviour.

We came across types of observational research in Year 1. Content analysis is a form of observation which analyses the communication that people produce. Anything from a single email or text to a series of films or television programmes may be an appropriate object of study.

KEY TERMS

Case studies – An in-depth investigation, description and analysis of a single individual, group, institution or event.

Content analysis – A research technique that enables the indirect study of behaviour by examining communications that people produce, for example, in texts, emails, TV, film and other media.

Coding – The stage of a content analysis in which the communication to be studied is analysed by identifying each instance of the chosen categories (which may be words, sentences, phrases, etc.).

Thematic analysis – An inductive and qualitative approach to analysis that involves identifying implicit or explicit ideas within the data. Themes will often emerge once the data has been coded.

Apply it

Concepts: Gynotikolobomassophobia

Patient X is a *gynotikolobomassophobic* – he has a morbid fear of women's ear lobes. His fear is so extreme that Patient X finds it impossible to talk to women in social situations (unless their ears are covered) and spends much of his time alone in his home.

A psychologist carrying out a case study of Patient X has conducted detailed interviews with him about his childhood. Patient X has also been encouraged to keep a diary as a record of his everyday experiences. The psychologist has concluded that Patient X's phobia may have been the result of childhood trauma.

Questions

1. What are the main features of a case study? Refer to Patient X as part of your answer.
2. Briefly discuss the strengths and limitations of the case study approach. Again, refer to Patient X as part of your discussion.
3. What ethical issues are associated with the case study approach?

Case studies

To study a 'case' in psychology is to provide a detailed and in-depth analysis of an individual, group, institution or event. **Case studies** often involve analysis of *unusual* individuals or events, such as a person with a rare disorder or the sequence of events that led to the 2011 London riots (see below). However, case studies may also concentrate on more 'typical' cases, such as an elderly person's recollections of their childhood.

Conducting a case study usually – though not exclusively – involves the production of **qualitative data**. Researchers will construct a **case history** of the individual concerned, perhaps using interviews, observations, questionnaires, or a combination of all of these. It is even possible that the person may be subject to experimental or psychological testing to assess what they are (or are not) capable of, and this may produce **quantitative data**.

Case studies tend to take place over a long period of time (**longitudinal**) and may involve gathering additional data from family and friends of the individual as well as the person themselves.

Content analysis

Content analysis is a type of observational research in which people are studied *indirectly* via the communications they have produced. The forms of communication that may be subject to content analysis are wide-ranging and may include spoken interaction (such as a conversation or speech/presentation), written forms (such as texts or emails) or broader examples from the media (such as books, magazines, TV programmes or films). The aim is to summarise and describe this communication in a systematic way so overall conclusions can be drawn.

Coding and quantitative data

Coding is the initial stage of content analysis. Some data sets to be analysed may be extremely large (such as the transcripts of several dozen lengthy interviews) and so there is a need to *categorise* this information into meaningful units. This may involve simply counting up the number of times a particular word or phrase appears in the text to produce a form of quantitative data. For instance, newspaper reports may be analysed for the number of times derogatory terms for the mentally ill are used, such as 'crazy' or 'mad'. Another example would be TV adverts which may be examined to see how often men and women are depicted in 'professional roles' (at work) or 'familial roles' (at home) (which is similar to a study carried out by Furnham and Farragher (2000) – see page 164 for more details).

Thematic analysis and qualitative data

Content analysis may also involve generating qualitative data, one example of which is **thematic analysis**. The process of coding and the identification of themes are closely linked insofar as themes may only emerge once data has been coded. A *theme* in content analysis refers to any idea, explicit or implicit, that is *recurrent* – in other words, which keeps 'cropping up' as part of the communication being studied. These are likely to be more descriptive than the coding units described above. For instance, the mentally ill may be represented in newspapers as 'a threat to the wellbeing of our children' or as 'a drain on the resources of the NHS'. Such themes may then be developed into broader categories, such as 'control', 'stereotyping' or 'treatment' of the mentally ill.

Once the researcher is satisfied that the themes they have developed cover most aspects of the data they are analysing, they may collect a new set of data to test the **validity** of the themes and categories. Assuming these explain the new data adequately, the researcher will write up the final report, typically using direct quotes from the data to illustrate each theme.

A scene from the London riots in 2011. Psychologists were interested in this one-off event and what it could tell us about so-called 'mob' behaviour.

Evaluation

Strengths

Case studies are able to offer rich, detailed insights that may shed light on very unusual and atypical forms of behaviour. This may be preferred to the more 'superficial' forms of data that might be collected from, say, an experiment or questionnaire.

As well as this, case studies may contribute to our understanding of 'normal' functioning. For example, the case of HM was significant as it demonstrated 'normal' memory processing – the existence of separate stores in STM and LTM.

Case studies may generate hypotheses for future study and one solitary, contradictory instance may lead to the revision of an entire theory – 'the single pebble that starts an avalanche'.

Limitations

Generalisation of findings is obviously an issue when dealing with such small sample sizes. Furthermore, the information that makes it into the final report is based on the subjective selection and interpretation of the researcher. Add to this the fact that personal accounts from the participants and their family and friends may be prone to inaccuracy and memory decay, especially if childhood stories are being told. This means that the evidence from case studies begins to look more than a little low in validity.

Evaluation

Strengths

Content analysis is useful in that it can circumnavigate (a posh word for 'get around') many of the **ethical issues** normally associated with psychological research. Much of the material that an analyst might want to study, such as TV adverts, films, personal ads in the newspaper or on the Internet, etc., may already exist within the public domain. Thus there are no issues with obtaining permission, for example. Communication of a more 'dubious' and sensitive nature, such as a conversation by text, still has the benefit of being high in **external validity**, provided the 'authors' consent to its use.

We have also seen that content analysis is flexible in the sense that it may produce both qualitative and quantitative data depending on the aims of the research.

Limitations

People tend to be studied *indirectly* as part of content analysis so the communication they produce is usually analysed *outside* of the context within which it occurred. There is a danger (similar to case studies above) that the researcher may attribute opinions and motivations to the speaker or writer that were not intended originally.

To be fair, many modern analysts are clear about how their own biases and preconceptions influence the research process, and often make reference to these as part of their final report (see the idea of **reflexivity** on page 95). However, content analysis may still suffer from a lack of objectivity, especially when more descriptive forms of thematic analysis are employed.

Apply it **Methods:**

Analysing driving behaviour

A researcher was interested to know whether there is a gender difference in driving behaviour and decided to conduct a **content analysis** of film clips of male and female drivers.

Question

Explain how the researcher might have carried out content analysis to analyse the film clips of driver behaviour. *(4 marks)*

Apply it **Concepts: Toilet humour**

Several studies in psychology have involved qualitative analysis of the content of *latrinalia* – that is, the graffiti often seen scribbled on toilet walls.

A more recent study by Matthews *et al.* (2012) involved the analysis of 1,200 instances of graffiti gathered from toilet walls in US bars. Graffiti was coded according to a number of distinct categories: *sexual references, socio-political (religion, politics, race, etc.), entertainment (music, TV), physical presence (the writing of one's name for instance), love/romance* and *scatological (for example, reference to defecation).* Graffiti was also classified in terms of whether it was *interactive* (a response to other graffiti) or *independent* (a stand-alone comment).

Matthews *et al.* found that males composed significantly more sexual and physical presence graffiti, whilst females authored more romantic and interactive graffiti.

Question

Explain how this investigation illustrates some of the strengths and limitations of content analysis.

Bathroom banter... but how might a content analysis of toilet wall graffiti be conducted?

Apply it **Methods:**

How to conduct a content analysis

Content analysis, like any observational research, involves design decisions about the following:

- *Sampling method* – how material should be sampled, e.g. **time sampling** or **event sampling**.
- *Recording data* – should data be transcribed or recorded, for instance, using video? Should data be collected by an individual researcher or within a team? (See the next spread for a discussion of the importance of **inter-rater reliability** when conducting content analysis.)
- *Analysing and representing data* – how should material be categorised or coded in order to summarise it? Should the number of times something is mentioned be calculated (quantitative analysis) or described using themes (qualitative analysis)?

Question

Explain how, in designing their study of latrinalia, Matthews *et al.* might have addressed each of the design decisions outlined above. *(6 marks)*

CHECK IT

1. Briefly evaluate the use of case studies in psychology. *[5 marks]*
2. Explain **one** limitation of using content analysis with research data. *[3 marks]*
3. Explain the processes involved in content analysis with reference to coding and thematic analysis. *[4 marks]*

RELIABILITY

THE SPECIFICATION SAYS...

Reliability across all methods of investigation. Ways of assessing reliability: test–retest and inter-observer; improving reliability.

In everyday life, when we describe someone (or something) as 'reliable', we mean that they are *dependable*; that we know to expect the same level of behaviour from them every single time. A reliable individual, for instance, is always punctual and never late or always late and never punctual. A reliable car is one that rarely breaks down and maintains the same level of performance over time.

Psychology's version of reliability is pretty similar: to what extent are the tests, scales, surveys, observations or experiments that psychologists use consistent – in the sense that their measurements of behaviour are the same (or at least similar) every single time they are used.

KEY TERMS

Reliability – Refers to how consistent the findings from an investigation or measuring device are. A measuring device is said to be reliable if it produces consistent results every time it is used.

Test–retest reliability – A method of assessing the reliability of a questionnaire or psychological test by assessing the same person on two separate occasions. This shows to what extent the test (or other measure) produces the same answers i.e. is consistent or reliable.

Inter-observer reliability – The extent to which there is agreement between two or more observers involved in observations of a behaviour. This is measured by correlating the observations of two or more observers. A general rule is that if (total number of agreements) / (total number of observations) > +.80, the data have high inter-observer reliability.

Reliability: it ain't great unless it's . . .

+.8

Statisticians don't write correlations with a leading zero and in reality they always write it as two decimal places but +.80 kinda spoils the rhyme!

Practical activity on page 84

Reliability

Reliability is a measure of *consistency*. In general terms, if a particular measurement can be repeated then that measurement is described as being reliable.

A ruler should find the same measurement for a particular object (let's say a chair) every time that object is measured – unless the ruler is broken or, in the words of Phoebe Buffay (*Friends*, Season 5, Episode 3), 'all the rulers are wrong'. If there is a change in the measurement over time, then we would attribute that change to the object rather than the ruler (someone may have sat on the chair and squashed it).

Similarly, if a test or measure in psychology assessed some 'thing' on a particular day (let's say intelligence), then we would expect the same result on a different day, unless the 'thing' itself had changed. Maybe we tested a different person with a different IQ or the same person's IQ went up a little (or possibly down after watching *Friends* . . .).

Unlike rulers, psychologists tend not to measure concrete things, like length or height, but are more interested in abstract concepts such as attitudes, aggression, memory and IQ. Can researchers have the same confidence in their **psychological tests**, **observations** and **questionnaires** as most of us—apart from Phoebe that is—have in a ruler?

Ways of assessing reliability

Test–retest

Psychologists have devised ways of assessing whether their measuring tools are reliable. The most straightforward way of checking reliability is the **test–retest** method. This simply involves administering the same test or questionnaire to the same person (or people) on different occasions. If the test or questionnaire is reliable then the results obtained should be the same, or at least very similar, each time they are administered. Note that this method is most commonly used with questionnaires and psychological tests (such as IQ tests) but can also be applied to **interviews**.

There must be sufficient time between test and retest to ensure, say, that the participant/respondent cannot recall their answers to the questions to a survey but not so long that their attitudes, opinions or abilities may have changed. In the case of a questionnaire or test, the two sets of scores would be **correlated** to make sure they are similar (see below). If the correlation turns out to be **significant** (and positive) then the reliability of the measuring instrument is assumed to be good..

Inter-observer reliability

The phrase '*beauty is in the eye of the beholder*' suggests that everyone has their own unique way of seeing the world. This issue is relevant to **observational research** as one researcher's interpretation of events may differ widely from someone else's – introducing **subjectivity**, **bias** and unreliability into the data collection process.

The recommendation is that would-be observers should not 'go it alone' but instead conduct their observations in teams of at least two. However, **inter-observer reliability** must be established. This may involve a small-scale trial run (a **pilot study**) of the observation in order to check that observers are applying **behavioural categories** in the same way, or it may be reported at the end of a study to show that the data collected was reliable. Observers obviously need to watch the same event, or sequence of events, but record their data independently. As with the test–retest method, the data collected by the two observers should be correlated to assess its reliability. Note that similar methods would apply to other forms of observation, such as **content analysis** (though this would be referred to as **inter-rater reliability**) as well as **interviews** if they are to be conducted by different people (known as **inter-interviewer reliability** – which is a bit of a mouthful).

Apply it — Concepts: The correlation 'test'

When assessing test–retest reliability or inter-observer reliability two sets of data will be correlated to see whether they match. The degree of correlation can be measured statistically using a **statistical test** of correlation such as **Spearman's rho** (see page 78).

Once the test has been performed on the two sets of data, a **correlation coefficient** will be calculated. The value of the coefficient must be +.80 or above for data to be judged reliable. Any figure lower than this and researchers must 'go back to the drawing board' so to speak and redesign their test or questionnaire – or reassess their observational categories.

Question

What would a correlation coefficient of +.95 between the data of two observers suggest?

Improving reliability

Questionnaires

As we have seen, the reliability of questionnaires over time should be measured using the test–retest method. Comparing two sets of data should produce a correlation that exceeds +.80 (see facing page). A questionnaire that produces low test–retest reliability may require some of the items to be 'deselected' or rewritten. For example, if some questions are complex or ambiguous, they may be interpreted differently by the same person on different occasions. One solution might be to replace some of the open questions (where there may be more room for (mis)interpretation) with closed, fixed choice alternatives which may be less ambiguous.

Interviews

For interviews, probably the best way of ensuring reliability is to use the same interviewer each time. If this is not possible or practical, all interviewers must be properly trained so, for example, one particular interviewer is not asking questions that are too **leading** or ambiguous. This is more easily avoided in **structured interviews** where the interviewer's behaviour is more controlled by the fixed questions. Interviews that are unstructured and more 'free-flowing' are less likely to be reliable.

Experiments

Lab experiments are often described as being 'reliable' because the researcher can exert strict control over many aspects of the procedure, such as the instructions that participants receive and the conditions within which they are tested. Certainly such control is often more achievable in a lab than in the field. This is more about precise **replication** of a particular *method* though rather than demonstrating the reliability of a *finding*. That said, one thing that may affect the reliability of a finding is if participants were tested under slightly different conditions each time they were tested.

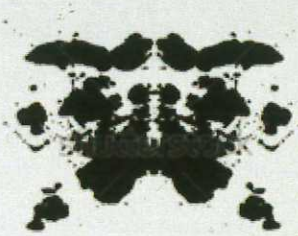
Observations

The reliability of observations can be improved by making sure that behavioural categories have been properly **operationalised**, and that they are measurable and self-evident (for instance, the category 'pushing' is much less open to interpretation than 'aggression'). Categories should not overlap ('hugging' and 'cuddling' for instance) and all possible behaviours should be covered on the checklist.

If categories are not operationalised well, or are overlapping or absent, different observers have to make their own judgements of what to record where and may well end up with differing and inconsistent records.

Apply it — Concepts: Personality testing

Personality tests in psychology take several forms and are often used in forensic settings to support clinical diagnosis (see the *Eysenck Personality Inventory (EPI)* on page 330). A more controversial measure of personality is the *Rorschach 'inkblot' test.* People are presented with a series of ambiguous inkblot images and are required to 'say what they see' in the pictures. The aim is to reveal the respondent's unconscious motivations and wishes as interpreted by the researcher or therapist. One criticism of the inkblot method is that one 'scorer' may not necessarily produce the same interpretation as another.



Questions

1. The inkblot-test has been criticised by many as an 'unreliable' measure of personality. Why do you think this is?
2. Explain *one* way of assessing the reliability of the Rorschach inkblot-test.

Q: What's the same as half an apple pie?

A: The other half!

Hilarious. But with halves of apple pie at least, we can assume reliability.

Apply it — Concepts: Inter-observer reliability amongst *Friends*

Two psychology students decided to see whether they could establish inter-observer reliability between themselves. They watched five episodes of *Friends* and recorded the different types of 'humour' within the programme. Before the study, they agreed on five observational categories of humour: sarcastic, slapstick, sexual/relationship-based, play on words and teasing.

Questions

1. Invent some data for their observations and put the data in a table. (*3 marks*)
2. The students compared their data and found a correlation coefficient of +.64, what does this indicate in terms of for the reliability of the two students' data? (*2 marks*)
3. What should the students do next to improve the reliability of their observation? (*4 marks*)

Apply it — Methods: Ghostly goings on – Part 1

A psychologist wanted to investigate the extent to which people believe in ghosts and devised a questionnaire as a way of assessing this. There were 20 items on the questionnaire in total.

Questions

1. Outline *one* way in which the psychologist could have assessed the **reliability** of the questionnaire. (*3 marks*)

Following the questionnaire, the psychologist selected a sample of 10 respondents who had completed the questionnaire and then observed their behaviour overnight in a house that was supposedly haunted. Working alongside another observer, the psychologist recorded evidence of a fear reaction to a number of stimuli including a creaking door, a gust of wind and a squeaky floorboard.

Questions

2. State *three* behavioural categories that could be used to measure the variable 'fear'. (*3 marks*)
3. Explain *one* way in which the researchers could have assessed the reliability of their observations. (*3 marks*)

CHECK IT

1. Outline what is meant by *reliability* in psychological research. *[2 marks]*
2. Explain **two** ways of assessing reliability. *[6 marks]*
3. Explain ways of improving reliability. *[5 marks]*

TYPES OF VALIDITY

THE SPECIFICATION SAYS...

Types of validity across all methods of investigation: face validity, concurrent validity, ecological validity and temporal validity. Assessment of validity. Improving validity.

Consistency within psychological research is one thing – but it is not the *only* thing. Demonstrating the same (or similar) findings on a number of different occasions is all very well – but what if the thing we are demonstrating each time turns out to be meaningless? Or not what we thought we were demonstrating? This is the issue of validity in psychological research – whether a study, investigation or investigative tool is a legitimate or genuine measure.

KEY TERMS

Validity – The extent to which an observed effect is genuine – does it measure what is was supposed to measure, and can it be generalised beyond the research setting within which it was found?

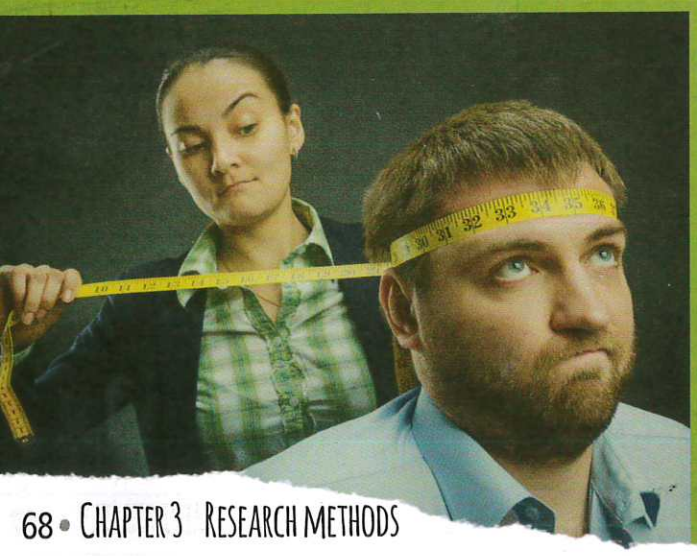
Face validity – A basic form of validity in which a measure is scrutinised to determine whether it appears to measure what it is supposed to measure – for instance, does a test of anxiety look like it measures anxiety?

Concurrent validity – The extent to which a psychological measure relates to an existing similar measure.

Ecological validity – The extent to which findings from a research study can be generalised to other settings and siutations. A form of external validity.

Temporal validity – The extent to which findings from a research study can be generalised to other historical times and eras. A form of external validity.

Whilst measuring your head produces a **reliable** result – in that it is the same (or similar) every time – as a measure of intelligence it is not **valid**.



Validity

Introducing validity

Validity refers to whether a **psychological test**, **observation**, **experiment**, etc., produces a result that is legitimate. In other words, whether the observed effect is genuine and represents what is actually 'out there' in the real world. This includes whether the researcher has managed to measure what they intended to measure (**internal validity**). It also refers the extent to which findings can be generalised beyond the research setting in which they were found (**external validity**).

It is possible for studies and measures to produce **reliable** data that is not valid. For instance, a broken set of scales may give a consistent reading of someone's weight which is always 7lbs more than their actual weight. In this example, the scales are reliable but the weight that is reported is not 'true' so the measurement lacks validity. In psychology, a test that claims to measure intelligence (or IQ) may not measure something 'true' about intelligence – it may simply measure a person's familiarity with IQ tests!

Internal validity

Internal validity refers to whether the effects observed in an experiment are due to the manipulation of the **independent variable** and not some other factor. One major threat to the internal validity of a study is if participants respond to **demand characteristics** and act in a way that they think is expected. For example, some commentators have questioned the internal validity of Milgram's obedience study claiming that participants were 'playing along' with the experimental situation and did not really believe they were administering shocks, i.e. they responded to the *demands* of the situation.

External validity

Meanwhile, external validity relates more to factors outside of the investigation, such as generalising to other settings, other populations of people and other eras.

Ecological validity

Ecological validity concerns generalising findings from one setting to other settings – most particular to 'everyday life' as that is what psychologists are interested in studying.

The concept of ecological validity is often misunderstood because students think it is about the naturalness of a study – a more natural setting should mean the findings from the study can be generalised to everyday life (high ecological validity). A lab is an artificial setting and therefore results of lab research should have low ecological validity because people don't behave naturally in a lab.

However, this isn't quite true. If the task that is used to measure the **dependent variable** in an experiment is not 'like everyday life' (i.e. low **mundane realism**) this can lower ecological validity. For example, a researcher might give people a list of words to remember to assess memory and could do this in a shopping mall – this would be a field study. However, in this case the *setting* doesn't make the findings more 'realistic'. The fact that we are using a word list makes the findings of the study lack ecological validity.

This means we must look at all sorts of aspects of the research set up in order to decide whether findings can be generalised beyond the particular research setting.

Temporal validity

Temporal validity is the issue of whether findings from a particular study, or concepts within a particular theory, hold true over time. Critics have suggested that high rates of conformity within the original Asch experiments were a product of a particularly conformist era in recent American history (the 1950s). Some of Freud's concepts, such as the idea that females experience **penis envy**, are deemed to be outdated, sexist and a reflection of the patriarchal Victorian society within which he lived.

STUDY TIP

Ecological validity versus mundane realism

*We have seen how the debate about whether findings from lab studies have ecological validity is often oversimplified. Both Asch's and Milgram's studies might be said to have high ecological validity as they were describing processes that often occur in everyday life (conformity and obedience). However, the tasks that participants had to complete within these studies (comparing line lengths and administering electric shocks) were not things people would normally be asked to do. Better to say then that the studies had low **mundane realism** as the experimental set-up did not mirror everyday life.*

Practical activity on page 84

Assessment of validity

One basic form of validity is **face validity**, whether a test, scale or measure appears 'on the face of it' to measure what it is supposed to measure. This can be determined by simply 'eyeballing' the measuring instrument or by passing it to an expert to check.

The **concurrent validity** of a particular test or scale is demonstrated when the results obtained are very close to, or match, those obtained on another recognised and well-established test. A new intelligence test, for instance, may be administered to a group of participants and the IQ scores they achieve may be compared with their performance on a well-established test (such as the *Stanford-Binet test*). Close agreement between the two sets of data would indicate that the new test has high concurrent validity – and close agreement is indicated if the correlation between the two sets of scores exceeds +.80.

Improving validity

Experimental research

In **experimental** research, validity is improved in many ways. For example, using a **control group** means that a researcher is better able to assess whether changes in the dependent variable were due to the effect of the **independent variable** (see Lombroso's research on page 326 for how the lack of a control group may affect validity). For instance, in a study looking at the effectiveness of a therapy, a control group who did not receive therapy means that the researcher can have greater confidence that improvement was due to effects of the therapy rather than, say, the passage of time.

Experimenters may also **standardise** procedures to minimise the impact of **participant reactivity** and **investigator effects** on the validity of the outcome. The use of **single-blind** and **double-blind procedures** is designed to achieve the same aim. You may remember that in a single-blind procedure participants are not made aware of the aims of a study until they have taken part (to reduce the effect of **demand characteristics** on their behaviour). In a double-blind study, a third party conducts the investigation without knowing its main purpose (which reduces both demand characteristics and investigator effects and thus improves validity).

Questionnaires

Many **questionnaires** and **psychological tests** incorporate a **lie scale** within the questions in order to assess the consistency of a respondent's response and to control for the effects of **social desirability bias**. Validity may be further enhanced by assuring respondents that all data submitted will remain **anonymous**.

Observations

Observational research may produce findings that have high ecological validity as there may be minimal intervention by the researcher. This is especially the case if the observer remains undetected, as in **covert observations**, meaning that the behaviour of those observed is likely to be natural and authentic.

In addition, **behavioural categories** that are too broad, overlapping or ambiguous may have a negative impact on the validity of the data collected.

Qualitative methods

Qualitative methods of research are usually thought of as having higher ecological validity than more **quantitative**, less interpretative methods of research. This is because the depth and detail associated with **case studies** and **interviews**, for instance, is better able to reflect the participant's reality.

However, the researcher may still have to demonstrate the **interpretive validity** of their conclusions. This is the extent to which the researcher's interpretation of events matches those of their participants. This can be demonstrated through such things as the *coherence* of the researcher's reporting and the inclusion of *direct quotes* from participants within the report. Validity is further enhanced through **triangulation** – the use of a number of different sources as evidence, for example, data compiled through interviews with friends and family, personal diaries, observations, etc.

Apply it — Methods: Ghostly goings on – Part 2

A psychologist wanted to investigate the extent to which people believe in ghosts and devised a questionnaire as a way of assessing this. There were 20 questions in total.

Questions

1. Explain what is meant by validity. Refer to the investigation above in your answer. (*3 marks*)
2. Explain *two* ways in which the psychologist could have improved the validity of the investigation above. (*4 marks*)

Apply it — Concepts: Threats to validity

The following are threats to validity that we came across as part of Research Methods in Year 1 – though some will apply to particular forms of research more than others.

Identify each from the definitions below:

1. Any variable, other than the IV, that may have an effect on the DV if it is not controlled. These are essentially nuisance variables that do not vary systematically with the IV.
2. Any variable, other than the IV, that may have affected the DV so we cannot be sure of the true source of changes to the DV. They vary systematically with the IV.
3. Any cue from the researcher or the research situation that may be interpreted by participants as revealing the true purpose of the investigation.
4. Any effect of the researcher's behaviour (conscious or unconscious) on the research outcome. This may include everything from the design of the study to the selection of, and interaction with, participants.
5. A question which, because of the way it is phrased, suggests a certain answer that may influence the response of the participant.

Did you get what you were aiming for? One of the concerns for psychologists trying to improve the validity of their research studies is that their expectations may influence the behaviour of their participants.

When assessing concurrent validity, the ***correlation coefficient*** between the two sets of scores must exceed +.80. Now where have we seen that before . . . ? It ain't great unless . . .

+.8

CHECK IT

1. Outline what is meant by *concurrent validity* in psychological research. *[2 marks]*
2. Distinguish between ecological validity and temporal validity as types of validity. *[6 marks]*
3. Explain **two** ways of assessing validity. *[6 marks]*
4. Explain ways of improving validity. *[5 marks]*

CHOOSING A STATISTICAL TEST

THE SPECIFICATION SAYS...

Factors affecting the choice of statistical test, including level of measurement and experimental design. When to use the following tests: Spearman's rho, Pearson's r, Wilcoxon, Mann–Whitney, related *t*-test, unrelated *t*-test and Chi-Squared test.

Quantitative (numerical) data can be summarised using descriptive statistics which include measures of central tendency, measures of dispersion, graphs and charts.

Although these are useful, they do not tell us whether the differences or correlations psychologists find are statistically significant (explained on the next spread), this is the job of statistical tests.

KEY TERMS

Levels of measurement – Quantitative data can be classified into types or levels of measurement, such as nominal, ordinal and interval.

Statistical tests – Used in psychology to determine whether a significant difference or correlation exists (and consequently, whether the null hypothesis should be rejected or retained).

Chi-Squared – A test for an association (difference or correlation) between two variables or conditions. Data should be nominal level using an unrelated (independent) design.

Mann–Whitney – A test for a significant difference between two sets of scores. Data should be at least ordinal level using an unrelated design (independent groups).

Pearson's r – A parametric test for correlation when data is at interval level.

Related *t*-test – A parametric test for difference between two sets of scores. Data must be interval with a related design, i.e. repeated measures or matched pairs.

Sign test – A statistical test used to analyse the difference in scores between related items (e.g. the same participant tested twice). Data should be nominal or better.

Spearman's rho – A test for correlation when data is at least ordinal level.

Unrelated *t*-test – A parametric test for difference between two sets of scores. Data must be interval with an unrelated design, i.e. independent groups.

Wilcoxon – A test for a significant difference between two sets of scores. Data should be at least ordinal level using a related design (repeated measures).

Choosing a statistical test

Statistical testing

In Year 1 you had a brief introduction to the concept of **statistical testing** using the example of the **sign test**. You will recall that a statistical test is used to determine whether a difference or an association found in a particular investigation is statistically **significant** – that is, more than could have occurred by **chance**. The outcome of this has implications for whether we accept or reject the **null hypothesis** – but we shall return to this shortly. For now, we need to consider which statistical test is used under what circumstances. There are three factors used to decide this:

1. Whether a researcher is looking for a *difference* or **correlation**.
2. In the case of a difference, what **experimental design** is being used.
3. The **level of measurement**.

These criteria are summarised in the table below.

1. Difference or correlation?

The first thing to consider when deciding which statistical test to use relates to the aim or purpose of the investigation – namely, is the researcher looking for a difference or correlation. This should be obvious from the wording of the **hypothesis**. In this context, 'correlation' can include *correlational analyses* as well as investigations that are looking for an *association* (see the **Chi-Squared** test on page 80).

2. Experimental design

You will also remember from Year 1 studies that there are three types of experimental design: **independent groups**, **repeated measures** and **matched pairs**. The last two of these are referred to as **related designs**. In a repeated measures design, the same participants are used in all conditions of the experiment. In a matched pairs design, participants in each condition are not the same but have been 'matched' on some variable that is important for the investigation which makes them 'related'. For this reason, both designs are classed as *related*.

As participants in each condition of an independent groups design are different, this design is **unrelated**. Thus, the researcher chooses from two alternatives here: *related* or *unrelated*.

Note that if the investigation is looking for a correlation, rather than a difference, then question 2 doesn't matter.

Choosing a statistical test

	Test of Difference		Test of association or correlation
	Unrelated design	Related design	
Nominal data	Chi-Squared	**Sign test**	Chi-Squared
Ordinal data	**Mann–Whitney**	**Wilcoxon**	**Spearman's rho**
Interval data	**Unrelated *t*-test**	**Related *t*-test**	**Pearson's r**

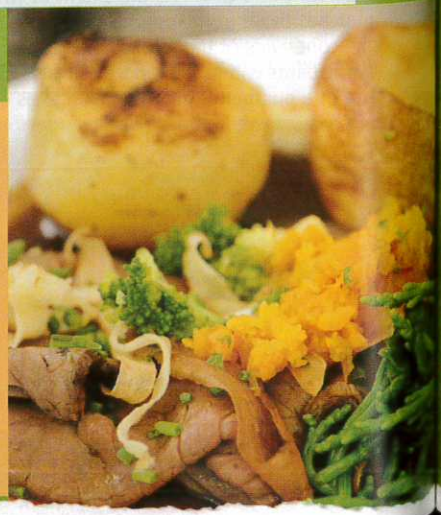
Note that Chi-Squared is a test of both difference and association/correlation. Data items must be unrelated.

Also note that the three tests on the blue background are parametric tests (the two forms of *t*-test and Pearson's r).

STUDY TIP

You will need to learn the table above so you know which test to use under what circumstances. If you are learning the table exactly as it looks here, the following mnemonic might help you remember the sequence of the tests (the first letter in each of the words in the sentence corresponds to the first letter of the stats test):

Carrots **S**hould **C**ome
Mashed **W**ith **S**wede
Under **R**oast **P**otatoes



3. Levels of measurement

Quantitative data can be divided into different **levels of measurement** and this is the third factor influencing the choice of statistical test. There are three levels of measurement: **nominal**, **ordinal** and **interval**.

Nominal data Data is represented in the form of categories – hence nominal data is sometimes referred to as **categorical data**. For example, you can count how many boys and girls in your Year group – male and female are the categories and you take a count of how many in each group.

Nominal data is **discrete** in that one item can only appear in one of the categories. For example, if you asked people to name their favourite football team their vote only appears in one category.

Ordinal data is ordered in some way. An example of ordinal data would be asking everyone in your class to rate how much they like psychology on a scale of 1 to 10 where 1 is 'do not like psychology at all' and 10 is 'absolutely love psychology'.

Ordinal data does not have equal intervals between each unit (unlike in interval data, below). For instance, in our example it would not make sense to say that someone who rated psychology an 8 enjoys it twice as much as someone who gave it a 4 .

Ordinal data also lacks precision because it is based on subjective opinion rather than objective measures. In our example, what constitutes a '4' or an '8' for the people doing the rating may be quite different. In the case of an IQ test the questions are derived from a view of what constitutes intelligence rather than any universal measurement. Questionnaires, psychological tests and so on do not measure something 'real' (i.e. they are not observable physical entities whereas, for example, reaction times and height are 'real'). Questionnaires etc. measure psychological constructs.

For these reasons, ordinal data is sometimes referred to as 'unsafe' data because it lacks precision. Due to its unsafe nature, ordinal data is not used as part of statistical testing. Instead, raw scores are converted to ranks (i.e. 1st, 2nd, 3rd, etc) and it is the ranks – not the scores – that are used in the calculation (see pages 74–75 and 78 for tests using ordinal data).

Interval data In contrast to ordinal data above, interval data is based on numerical scales that include units of equal, precisely defined size. In this sense it is 'better' than ordinal data because more detail is preserved (and ordinal is 'better' than nominal level).

Think of the kinds of things you would use to take measurements with in maths or other sciences, such as a stopwatch, a thermometer or weighing scales. These are public scales of measurement that produce data based on accepted units of measurement (time, temperature, weight). So, for instance, if we recorded how long it took each of our students to complete a written recall test in psychology, we would have collected interval data. Interval data is the most precise and sophisticated form of data in psychology and is a necessary criterion for the use of parametric tests (see right).

Table showing levels of measurement and their relation to the appropriate measures of central tendency and measures of dispersion.

Level of measurement	Measure of central tendency	Measure of dispersion
Nominal	Mode	n/a
Ordinal	Median	Range
Interval	Mean	Standard deviation

Note that the range and standard deviation cannot be calculated on nominal data as such data is in the form of frequencies. It is not appropriate to use the mean or the standard deviation for ordinal data as the intervals between the units of measurement are not of equal size.

Apply it Concepts: Which level of measurement?

Identify whether the following would produce data at the nominal, ordinal or interval level.

1. Time taken to sort cards into categories.
2. Peoples' choice of the *Sun*, *The Times* or the *Guardian*.
3. Participants' sense of self-worth, estimated on a scale of 1–10.
4. Judges in a dancing competition giving marks for style and presentation.
5. Participants' reaction to aversive stimuli measured using a heart rate monitor.
6. A set of medical records classifying patients as either chronic, acute or 'not yet classified'.

STUDY TIPS

Some of the data produced in psychology is quite difficult to classify. For example, should we treat 'number of words recalled' in a memory test as interval or ordinal data?

Strictly speaking, this would only be interval data if the words are all of equal difficulty (so the units of measurement are all equivalent). This would be very difficult to achieve as some words will always be more memorable than others! For this reason, it is probably 'safer' to treat number of words recalled as ordinal data and rank the set of scores accordingly.

But you should always provide your reasoning when deciding which level of measurement is appropriate.

Apply it Concepts: Parametric tests

The related *t*-test, unrelated *t*-test and Pearson's r are collectively known as **parametric tests**. Parametric tests are more powerful and robust than other tests. If a researcher is able to use a parametric test they will do so, as these tests may be able to detect significance within some data sets that non-parametric tests cannot.

There are three criteria that must be met in order to use a parametric test:

1. Data must be **interval level** – parametric tests use the actual scores rather than ranked data.
2. The data should be drawn from a population which would be expected to show a **normal distribution** for the variable being measured. Variables that would produce a **skewed distribution** are not appropriate for parametric tests.
3. There should be *homogeneity of variance* – the set of scores in each condition should have similar dispersion or spread. One way of determining variance is by comparing the standard deviations in each condition; if they are similar, a parametric test may be used. In a related design it is generally assumed that the two groups of scores have a similar spread.

Question

If a researcher compared two related sets of data and was looking to see if they were different, why would it be preferable to use a related *t*-test instead of a Wilcoxon?

CHECK IT

1. Identify and explain the difference between **two** levels of measurement in psychological research. *[4 marks]*
2. Identify **three** factors that influence the choice of statistical test. *[3 marks]*
3. Explain **two** factors that would be required for use of an unrelated *t*-test. *[4 marks]*

PROBABILITY AND SIGNIFICANCE

THE SPECIFICATION SAYS...

Probability and significance: use of statistical tables and critical values in interpretation of significance; Type I and Type II errors.

All statistical tests end with a number – the calculated value. This number is crucial in determining whether the researcher has found a result that is statistically significant, and consequently, whether they should accept the alternative or null hypothesis.

To understand how statistical tests work requires an understanding of the related concepts of probability and significance.

KEY TERMS

Probability – A measure of the likelihood that a particular event will occur where 0 indicates statistical impossibility and 1 statistical certainty.

Significance – A statistical term that tells us how sure we are that a difference or correlation exists. A 'significant' result means that the researcher can reject the null hypothesis.

Critical value – When testing a hypothesis, the numerical boundary or cut-off point between acceptance and rejection of the null hypothesis.

Type I error – The incorrect rejection of a true null hypothesis (a false positive).

Type II error – The failure to reject a false null hypothesis (a false negative).

What is the probability of two people in a football match sharing the same birthday? There are 23 people on the pitch (including the referee). The chance that any two people will have the same birthday is 1 in 365. If all 23 people shook hands with each other, there would be 253 handshakes. This equates to the number of pairs of people who could potentially share the same birthday. 253/365 = 0.69. The probability of two people in a football match sharing the same birthday is 69% i.e. well over half. Most people are surprised by how high this is!

Probability and significance

The null hypothesis

Researchers begin their investigations by writing a **hypothesis**. This may be **directional** or **non-directional** depending how confident the researcher is in the outcome of the investigation. Here is a an example of a hypothesis (you may remember it from the Year 1 book):

After drinking 300ml of SpeedUpp, participants say more words in the next five minutes than participants who drink 300ml of water.

This is sometimes referred to as an **alternative hypothesis** (or H_1 for short) because it is alternative to the **null hypothesis** (H_0). The null hypothesis states there is 'no difference' between the conditions:

There is no difference in the number of words spoken in five minutes between participants who drink 300ml of SpeedUpp and participants who drink 300ml of water.

The **statistical test** determines which hypothesis is 'true' and thus whether we accept or reject the null hypothesis.

Levels of significance and probability

Actually, 'true' is probably the wrong word. Statistical tests work on the basis of **probability** rather than certainty. All statistical tests employ a **significance level** – the point at which the researcher can claim to have discovered a significant difference or correlation within the data. In other words, the point at which the researcher can reject the null hypothesis and accept the alternative hypothesis.

*The usual level of **significance** in psychology is 0.05 (or 5%).*

This is properly written as $p \leq 0.05$ (where p stands for probability).

This means the probability that the observed effect (the result) occurred by chance is equal to or less than 5%. In effect, this means that even when a researcher claims to have found a significant difference/correlation, there is still up to 5% probability that the observed effect occurred by chance – that it was a 'fluke'.

Psychologists can never be 100% certain about a particular result as they have not tested all members of the population under all possible circumstances! For this reason, psychologists have settled upon a conventional level of probability where they are prepared to accept that results may have occurred by chance – this is the 5% level.

STUDY TIPS

People often refer to probability and chance in everyday life. We might surmise that the chance of rain is around '50/50', that our favourite football team has a 'good chance' of winning on Saturday or that we have 'no chance' of winning the National Lottery (the actual statistical probability is about 1 in 14 million).

In psychological research, the 5% significance level ensures that, in the case of a significant result, there is equal to or less than 5% probability that the result occurred by chance. However, in these circumstances, it is not correct to state that we can be '95% certain that the result did not occur by chance'. If you think about it, the phrase '95% certain' is a contradiction in terms – we can only ever be 100% certain of anything – and statistical testing deals with probabilities not certainties!

Apply it — Concepts: Drug testing

A researcher is testing the effectiveness of a new drug that relieves the symptoms of anxiety disorder – *Anxocalm*. The researcher is comparing two groups of people who suffer from anxiety: one group will complete a course of Anxocalm and the other group will be given a **placebo**. There is a possibility that the drug may cause mild side effects in those who take it (such as a headache and nausea). For this reason, the researcher can only test the drug once on human participants.

The researcher has decided to use the 1% level when testing for significance.

Question

Explain why the researcher has decided to use the 1% level of significance on this occasion.

Use of statistical tables

The critical value

Once a statistical test has been calculated, the result is a *number* – the **calculated value** (or observed value). To check for statistical significance, the calculated value must be compared with a **critical value** – a number that tells us whether or not we can reject the null hypothesis and accept the alternative hypothesis.

Each statistical test has its own **table of critical values,** developed by statisticians. These tables look like very complicated bingo cards (see example on the next spread). For some statistical tests, the calculated value must be equal to or greater than the critical value; for other tests, the calculated value must be equal to or less than the critical value (see the 'Rule of R' below).

Using tables of critical values

How does the researcher know which critical value to use? There are three criteria:

- **One-tailed** or **two-tailed test**? You use a one-tailed test if your hypothesis was directional and a two-tailed test for a non-directional hypothesis. Probability levels *double* when two-tailed tests are being used as they are a more *conservative* prediction.
- The number of participants in the study. This usually appears as the *N* value on the table. For some tests **degrees of freedom (df)** are calculated instead.
- The **level of significance** (or *p* value). As discussed, the 0.05 level of significance is the standard level in psychological research.

Lower levels of significance

Occasionally, a more stringent level of significance may be used (such as 0.01) in studies where they may be a *human cost* – such as drug trials – or 'one-off' studies that could not, for practical reasons, be repeated in future. In all research, if there is a *large* difference between the calculated and critical values – in the preferred direction – the researcher will check more stringent levels, as the *lower* the p value is, the more statistically significant the result.

Type I and Type II errors

Due to the fact that researchers can never be 100% certain that they have found statistical significance, it is possible (*usually up to 5% possible*) that the wrong hypothesis may be accepted.

A **Type I error** is when the null hypothesis is rejected and the alternative hypothesis is accepted when it should have been the other way round because, in reality, the null hypothesis is 'true'. This is often referred to as an optimistic error or false positive as the researcher claims to have found a significant difference or correlation when one does not exist.

A **Type II error** is the reverse of the above: when the null hypothesis is accepted but it should have been the alternative hypothesis because, in reality, the alternative hypothesis is true. This is a pessimistic error or 'false negative'.

We are more likely to make a Type I error if the significance level is too lenient (too high) e.g. 0.1 or 10% rather than 5%. A Type II error is more likely if the significance level is too stringent (too low) e.g. 0.01 or 1%, as potentially significant values may be missed. Psychologists favour the 5% level of significance as it best balances the risk of making a Type I or Type II error.

Apply it **Concepts: Pregnancy tests**

Pregnancy tests are not 100% reliable so women who suspect they are pregnant are advised to take more than one test in order to make sure.

Question

If the result says you are not pregnant – in what way could this be a Type II error?

STUDY TIP

If you are testing a directional hypothesis you may find that your calculated value is significant – but there is a further issue. Are your results in the direction you predicted? If they are not, then you must accept the null hypothesis even though the calculated value is significant. Before you ask – you can't just change the original hypothesis!

In fact, in such cases this should be obvious when looking at the data and a researcher would not carry out any statistical testing.

STUDY TIP

As suggested above, it is okay to check more stringent levels of significance as long as the critical value at the 5% level has been checked first to establish significance. However, higher levels of significance, such as 10% should be disregarded. At these levels, the null hypothesis cannot be rejected – though the hypothesis may be worth pursuing and refining the methodology.

Apply it **Concepts: The rule of R**

Some statistical tests require the calculated value to be *equal to or more than* the critical value for statistical significance; for other tests, the calculated value must be *equal to or less than* the critical value.

The *rule of R* can help with this. Those statistical tests with a letter 'R' in their name are those where the calculated value must be equal to or *more* than the critical value (note that there is also an 'r' in 'more' which is a further clue!)

Questions

1. List the statistical tests with a letter R in their name.
2. List the statistical tests without a letter R.

CHECK IT

1. Distinguish between a type I and type II error in psychological research. *[3 marks]*
2. Define what is meant by the *critical value* in statistical testing. *[2 marks]*
3. What is the accepted level of significance in psychological research? *[1 mark]*

Tests of difference: Mann–Whitney and Wilcoxon

The specification says...

Students should be familiar with the use of inferential tests.

An 'inferential test' is another term for a statistical test. In Year 1 of the course you learned to use a statistical test of difference – the sign test. This spread includes two further statistical tests that are used to determine whether two samples are significantly different: Mann–Whitney and Wilcoxon. In each case a worked example is given so you can understand how the test is calculated and how significance is determined.

Note that, in an independent groups design, the numbers of participants in each group may be different as is the case here – Group A has 10 participants and group B has 8 participants.

Table 3 Critical values of U for a two-tailed test, $p \leq 0.05$

N_A / N_B	2	3	4	5	6	7	8	9	10
2							0	0	0
3				0	1	1	2	2	3
4			0	1	2	3	4	4	5
5		0	1	2	3	5	6	7	8
6		1	2	3	5	6	8	10	11
7		1	3	5	6	8	10	12	14
8	0	2	4	6	8	10	13	15	17
9	0	2	4	7	10	12	15	17	20
10	0	3	5	8	11	14	17	20	23
11	0	3	6	9	13	16	19	23	26
12	1	4	7	11	14	18	22	26	29
13	1	4	8	12	16	20	24	28	33
14	1	5	9	13	17	22	26	31	36

Calculated value of U must be EQUAL TO or LESS THAN the critical value in this table for significance to be shown.

Apply it

Methods: What does it all mean...

The investigation described on the right found a significant difference at $p \leq 0.05$.

1. Explain what is meant by the phrase 'a significant difference was found at $p \leq 0.05$'. (*2 marks*)
2. What conclusion can be drawn from the investigation described? (*2 marks*)

Mann–Whitney: A worked example

Why Mann–Whitney?

In this worked example we are looking for a difference between two groups of employers based on their rating of whether a candidate (who had suffered from **schizophrenia**) was suitable for a job interview. There are two **independent groups** of employers, which means the design is unrelated. Finally, the level of measurement is **ordinal** as data is based on scores on an 'unsafe' scale (subjective ratings of interview suitability) which are converted to ranks for the purposes of the test.

The aim...

A study of the effects of labelling in schizophrenia was conducted to see if there is a difference in someone's perceived 'employability' based on whether they had been diagnosed with schizophrenia in the past. Eighteen employers were shown an application form and ask to rate the candidate in terms of how likely they would be called for an interview, on a scale of 1–20 (where 1 = definitely would not be interviewed and 20 = definitely would be interviewed).

All employers saw the same application form, the only difference was that for employers in Group A the form included the phrase 'recovering schizophrenic'. For employers in Group B, that phrase was absent from the form.

The hypotheses...

Alternative hypothesis: *There is a difference in ratings for 'suitability for an interview' based on whether a job applicant is described as having been diagnosed with schizophrenia (Group A) or not (Group B).* (**non-directional, two-tailed**)

Null hypothesis: *There is no difference in ratings for 'suitability for an interview' based on whether a job applicant is described as having been diagnosed with schizophrenia (Group A) or not (Group B).*

Step 1: The table of ranks...

To rank the ratings you need to consider the data from both Groups A and B at the same time (data is given in the table below). The lowest number has a rank of 1. In the case where two data items are the same you add up the rank they would get and give the mean for those ranks. For example the rating of 12 appears four times in the table at rank position 7, 8, 9 and 10 therefore they all are given the rank of 8.5.

Where there are a lot of multiple ranks it may help to use a frequency table (see Table 1).

Calculate the sum of the ranks for Group A (R_A) and for Group B (R_B) (see Table 2).

Table 1 Frequency table

Rating	*Frequency*	*Rank*
8	|	1
9	|	2
10	||	3 and 4
11	||	5 and 6
12	||||	7, 8, 9 and 10
13	|	11
14	|	12
15	||	13 and 14
16	|	15
17	||	16 and 17
18	|	18

Table 2 Calculations table

Group A participant number	*Suitability for interview rating*	*Rank*	*Group B participant number*	*Suitability for interview rating*	*Rank*
1	12	8.5	11	16	15
2	10	3.5	12	12	8.5
3	13	11	13	14	12
4	8	1	14	15	13.5
5	12	8.5	15	18	18
6	10	3.5	16	17	16.5
7	11	5.5	17	11	5.5
8	15	13.5	18	17	16.5
9	9	2			
10	12	8.5			
		R_A =65.5			R_B =105.5

Step 2: Working out the value of U...

Calculate the smaller value of U, which in this case will be Group A (the value of U is now called U_A and the number of participants in group A is referred to as N_A).

$U = U_A = R_A - [N_A (N_A + 1)] / 2 = 65.5 - [10 \times (10 + 1)] / 2 = 10.5$

Step 3: The calculated and critical values...

The **calculated value** of U is **10.5**

The **critical value** (in Table 3) of U for a two-tailed test at the 0.05 level where N_A =10 and N_B = 8 is 17 (see table of critical values, above left).

As the calculated value of U is less than the critical value the result is **significant** ($p \leq 0.05$) and we can reject the null hypothesis and accept the alternative hypothesis: *There is a difference in ratings for 'suitability for an interview' based on whether a job applicant is described as having been diagnosed with schizophrenia (Group A) or not (Group B)* ($p \leq 0.05$).

Practical activity on page 85

Wilcoxon: A worked example

Why Wilcoxon?

In this worked example we are looking for a difference in anger scores before and after using an **anger management** programme. This is a **repeated measures** design (i.e. related) as the same participants are assessed before and after receiving treatment. The data is **ordinal** as anger scores are based on a subjective 'unsafe' **self-report questionnaire**.

The aim...

An investigation in forensic psychology was conducted to assess the effectiveness of a new anger management programme. Twenty teenagers serving time in a young offenders institute for violent crime were involved in the study. At the beginning of the investigation, all the offenders completed a questionnaire to measure their level of anger. This gave each offender an anger score out of 50. The offenders then completed eight intensive sessions of anger management. Following the treatment, the offenders completed the same anger questionnaire. The two sets of scores – before and after treatment – were compared to see if there was a difference.

The hypotheses...

Alternative hypothesis: *There is a difference in young offenders' scores on an anger questionnaire before and after treatment.* (**non-directional**, **two-tailed**).

Null hypothesis: *There is no difference in young offenders' scores on an anger questionnaire before and after treatment.*

Step 1: Calculate a difference and rank the difference...

This time ranking is done on the difference between the two sets of data. When ranking, the signs are ignored.

If the difference is zero the data is not included in the ranking, as below.

Table 4 Calculations table

Participant	*Anger score before treatment*	*Anger score after treatment*	*Difference*	*Rank of difference*
1	39	30	+9	7.5
2	42	44	-2	1
3	28	25	+3	3
4	35	32	+3	3
5	32	32	-	-
6	40	30	+10	9
7	50	44	+6	6
8	46	50	-4	5
9	29	20	+9	7.5
10	44	29	+15	10
11	25	28	-3	3
12	38	38	-	-

Step 2: Working out the value of T

The **calculated value** of T is the sum of the less frequent sign. The less frequent sign is *minus*, so the sum of the ranks is 1 + 5 + 3.

T = 9

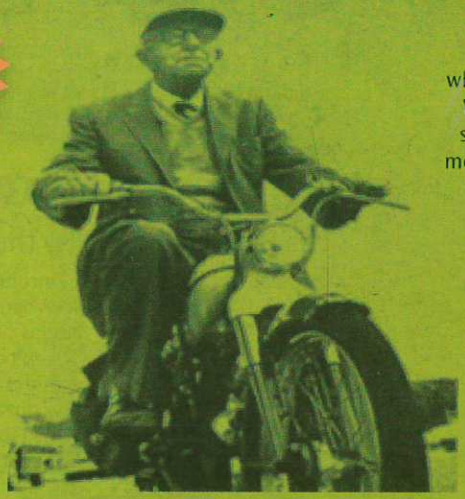
Step 3: The calculated and critical values...

The **calculated value** of T is 9.

The **critical value** of T for a two-tailed test at the 0.05 level when N = 10 is 8 (see table of critical values, above right).

As the calculated value of T is more than the critical value of T the result is not significant ($p \leq 0.05$) and we must accept the null hypothesis: *There is no difference in young offenders' scores on an anger questionnaire before and after treatment* ($p > 0.05$).

We reject the alternative hypothesis at $p \leq 0.05$ (i.e. less than a 5% probability that the results are due to chance) and therefore accept the null hypothesis at $p > 0.05$ (i.e. there was more than a 5% probability the results are due to chance).



Henry Wilcoxon, who brought us the Wilcoxon test (not surprisingly). On a motorbike (perhaps surprisingly).

Table 5 Critical values of T

Level of significance for a one-tailed test	0.05	0.025	0.01
Level of significance for a two-tailed test	0.10	0.05	0.02
N = 5	0		
6	2	0	
7	3	2	0
8	5	3	1
9	8	5	3
10	11	8	5
11	13	10	7
12	17	13	9
13	21	17	12
14	25	21	15
15	30	25	19

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

Apply it

Methods: Using the critical value table

In a similar investigation, a **matched pairs design** was used to assess the effectiveness of the anger management programme. 20 offenders were matched on anger score at the beginning of the investigation and one from each pair was allocated either to the treatment condition (eight sessions of anger management) or the control condition (no treatment). Anger scores were assessed at the end of the investigation.

The calculated value of T was 6. The hypothesis was non-directional. Note that, in a matched pairs design, the N value is based on the number of pairs (10).

Questions

1. Is the result significant? Explain your answer. (*3 marks*)
2. What conclusion can be drawn from this study? (*2 marks*)

CHECK IT

A researcher was interested to know whether there was a gender difference in 'enjoyment rating' of A level Psychology students.

1. Which statistical test would be used to analyse the data? Justify your choice. *[4 marks]*
2. When would a researcher decide to use a Wilcoxon test? Refer to **three** factors in your answer. *[3 marks]*

PARAMETRIC TESTS OF DIFFERENCE: UNRELATED AND RELATED T-TESTS

THE SPECIFICATION SAYS

Students should be familiar with the use of inferential tests.

There are two other difference tests that can be used when data is interval level instead of the less powerful non-parametric Mann–Whitney and Wilcoxon tests. These are the two tests of difference, the unrelated and related *t*-test.

Table 2 Critical values of t

Level of significance for a one-tailed test	0.05	0.025
Level of significance for a two-tailed test	0.10	0.05
df= 1	6.314	12.706
2	2.920	4.303
3	2.353	3.182
4	2.132	2.776
5	2.015	2.571
6	1.943	2.447
7	1.895	2.365
8	1.860	2.306
9	1.833	2.262
10	1.812	2.228
11	1.796	2.201
12	1.782	2.179
13	1.771	2.160
14	1.761	2.145
15	1.753	2.131
16	1.746	2.120
17	1.740	2.110
18	1.734	2.101
19	1.729	2.093
20	1.725	2.086
21	1.721	2.080
22	1.717	2.074
23	1.714	2.069
24	1.711	2.064
25	1.708	2.060
26	1.706	2.056
27	1.703	2.052
28	1.701	2.048
29	1.699	2.045
30	1.697	2.042
40	1.684	2.021
60	1.671	2.000
120	1.658	1.980

Calculated value of *t* must be EQUAL TO or MORE THAN the critical value in this table for significance to be shown.

Apply it

Methods: Increasing sample size

Question

If the same investigation was repeated with 61 boys and 61 girls, and the same calculated value was achieved, would the result be significant? Explain your answer. *(3 marks)*

Unrelated *t*-test: A worked example

Why the unrelated *t*-test?

The **unrelated *t*-test** is a test of difference between two sets of data. It is used with interval level data only. When an **independent groups design** is used, the test selected is the unrelated *t*-test.

In this worked example, we are looking for a difference in the time taken to complete a jigsaw puzzle between boys and girls. The type of design is independent groups (unrelated) because one group were girls and the other group were boys. The level of measurement is **interval** as time taken to complete a jigsaw puzzle is measured on a 'safe' scale (a scale of public measurement) made up of equal units. It is assumed that the participants are drawn from a **normally distributed sample** within the population and there is **homogeneity of variance** as the standard deviations in both groups are similar.

The aim ...

An investigation of **gender** looked into whether there was a difference in visuo-spatial ability between boys and girls. Ten girls and ten boys took part in the test which involved completing a simple jigsaw puzzle in the shortest time possible. All participants completed the same puzzle and the time it took for each of them was recorded and compared.

The hypotheses ...

Alternative hypothesis: *There is a difference in the time taken by males and females to complete a jigsaw puzzle.* (**non-directional, two-tailed**)

Null hypothesis: *There is no difference in the time taken by males and females to complete a jigsaw puzzle.*

Step 1: The table of data ...

In Table 1 below various calculations need to be made for the Group A and B scores:

- Calculate the sum of the scores for Group A (ΣX_A). (X_A refers to scores in group A.)
- Repeat for Group B (ΣX_B).
- Square each value in Group A (X_A^2).
- Repeat for Group B (X_B^2).

Σ means 'sum of'. See completed table (below).

Table 1 Calculations table

Group A male participants	*Time taken (sec) X_A*	X_A^2	*Group B female participants*	*Time taken (sec) X_B*	X_B^2
1	64	4096	1	52	2704
2	56	3136	2	59	3481
3	89	7921	3	90	8100
4	55	3025	4	112	12544
5	79	6241	5	84	7056
6	102	10404	6	73	5329
7	80	6400	7	79	6241
8	69	4761	8	64	4096
9	69	4761	9	49	2401
10	80	6400	10	90	8100
	$\Sigma X_A = 743$	$\Sigma X_A^2 = 57145$		$\Sigma X_B = 752$	$\Sigma X_B^2 = 60052$

Step 2: Working out the value of t ...

$$t = \frac{(\bar{X}_A - \bar{X}_B)}{\sqrt{\left(\frac{S_A + S_B}{N_A + N_B - 2}\right) \times \left(\frac{N_A + N_B}{N_A N_B}\right)}}$$

$\bar{X}$ stands for the mean.

N_A and N_B are the numbers of scores in group A and B.

Where: $S_A = \Sigma X_A^2 - (\Sigma X_A)^2 / N_A$

$S_B = \Sigma X_B^2 - (\Sigma X_B)^2 / N_B$

$S_A = 57145 - 55204.9 = 1940.1$

$S_B = 60052 - 56550.4 = 3501.6$

$$t = \frac{(74.3 - 75.2)}{\sqrt{\left(\frac{1940.1 + 3501.6}{10 + 10 - 2}\right) \times \left(\frac{10 + 10}{100}\right)}} = -0.116$$

Step 3: The calculated and critical values ...

The **calculated value** of t = −0.116 (note that t is a negative value because the mean for group B was larger than group A, when checking the critical values table ignore the negative sign).

The **critical value** (in Table 2) for a two-tailed test at the 0.05 level where $df = N_A + N_B - 2 = 18$, is 2.101.

As the calculated value (ignoring the sign) is less than the critical value ($p \leq 0.05$) the result is not significant and we must accept the null hypothesis: *There is no difference between males and females in the time taken to complete a jigsaw puzzle* ($p > 0.05$).

Practical activity on page 85

Related *t*-test: A worked example

Why the related *t*-test?

When a **repeated design** is used the test selected is the **related *t*-test**.

Here, we are looking for a difference in the average heart rate before and after treatment (**CBT**). The type of design is **repeated measures** (related) because the same participants were tested twice. The level of measurement is interval as measurements of heart rate (beats per minute, bpm) are based on a 'safe' scale (a scale of public measurement) made up of equal units. Let us assume for the purpose of this test that participants were drawn from a **normally distributed sample** within the population and **homogeneity of variance** is assumed as this is a related design.

The aim...

In a study of **addiction,** researchers investigated the effects of CBT on the physiological arousal of gamblers. Ten participants who were categorised as 'persistent gamblers' were given a six-week course of CBT to change their gambling behaviour. Before treatment, all of the participants played on a fruit machine for 20 minutes whilst their heart rate activity was monitored as a measure of physiological arousal. Following treatment, the same participants played on the same game for the same length of time and their heart rate activity was monitored.

The hypotheses...

Alternative hypothesis; *There is a reduction in heart rate activity when comparing heart rate before and after cognitive behaviour therapy.* **(directional, one-tailed)**

Null hypothesis: *There is no difference in heart rate activity comparing heart rate before and after cognitive behaviour therapy.*

Step 1: The table of data...

In the Table 3 below, various calculations need to be made for condition A and B: Calculate the difference between the two sets of scores (d). See completed Table 3.

- Calculate the difference (d) between scores for condition A and condition B.
- Square each difference (d^2).
- Add up the values in the d column to give the sum of d (Σd).
- Add up the values in the d^2 column to give the sum of d^2 (Σd^2).

Table 3 Calculations table

Participant	*Condition A Heart rate (bpm) before treatment*	*Condition B Heart rate (bpm) after treatment*	*Difference* (d)	d^2
1	84	80	4	16
2	71	70	1	1
3	52	55	−3	9
4	66	58	8	64
5	58	58	0	0
6	77	70	7	49
7	63	61	2	4
8	81	75	6	36
9	71	74	−3	9
10	70	61	9	81
			$\Sigma d = 31$	$\Sigma d^2 = 269$

Step 2: Working out the value of t...

$$t = \frac{(\Sigma D)/N}{\sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{N}}{N(N-1)}}} \qquad t = \frac{31/10}{\sqrt{\frac{269 - \frac{961}{10}}{10(10-1)}}} \qquad t = \frac{3.1}{\sqrt{\frac{172.9}{90}}} \qquad t = \frac{3.1}{1.386} \qquad = 2.236$$

Step 3: The calculated and critical values...

The **calculated value** of t is 2.236

The **critical value** of t (in Table 2 on the facing page) for a one-tailed test at the 0.05 level where *df* is N − 1 = 9, is 1.833

As the calculated value of t is greater than the critical value ($p > 0.05$) the result is significant and we can reject the null hypothesis and conclude: *There is a reduction in heart rate activity comparing heart rate before and after cognitive behaviour therapy* ($p > 0.05$).

Andrea appeared to have misunderstood the suggestion from her psychology teacher that she should carry out a *t*-test.

Apply it

Methods: *t*-tests and taxi drivers

Read the Maguire *et al.* taxi driver study on page 40. A different researcher wanted to assess whether there was a *change* in taxi drivers' hippocampal volume as a result of taking 'The Knowledge' test. They analysed the hippocampal volume of 30 trainee London cabbies before they began studying for the test. After all the drivers had completed their training and taken 'The Knowledge' test, the researchers took the same measurement again.

Questions

1. Write a directional hypothesis for the investigation described above. *(2 marks)*
2. Which of the two *t*-tests should be used to analyse the data? Justify your answer. *(2 marks)*

The researcher analysed the data. The calculated value of t was 1.526.

3. Is the result significant? Explain your answer. *(3 marks)*
4. What conclusion can be drawn from this study? *(2 marks)*

CHECK IT

A researcher wanted to know whether A level PE students could throw a ball further than A level Geography students.

1. Which statistical test would be used to analyse the data? Justify your choice. *[4 marks]*
2. When would a researcher decide to use a related *t*-test? Refer to **three** factors in your answer. *[3 marks]*

Tests of correlation: Spearman's and Pearson's

Practical activity for both tests on pages 84 and 85

THE SPECIFICATION SAYS...

Students should be familiar with the use of inferential tests.

And the statistical (inferential) tests keep on coming.... Both of the tests featured here – Spearman's rho and Pearson's r – are looking for a correlation between co-variables rather than a difference between sets of scores.

Spearman's can be used with *ordinal* or *interval* data. Pearson's test can only be used if the data are *interval*.

Table 2 Critical values of rho

Level of significance for a one-tailed test	0.05	0.025
Level of significance for a two-tailed test	0.10	0.05
N = 1	1.000	
5	.900	1.000
6	.829	.886
7	.714	.786
8	.643	.738
9	.600	.700
10	.564	.648
11	.536	.618
12	.503	.587
13	.484	.560
14	.464	.538
15	.443	.521
16	.429	.503
17	.414	.485
18	.401	.472
19	.391	.460
20	.380	.447
25	.337	.398
30	.306	.362

Calculated value of rho must be EQUAL TO or MORE THAN the critical value in this table for significance to be shown.

Apply it

Methods: Substituting and estimating values

A similar investigation with the same hypothesis was conducted with 21 couples. The sum of the difference between the ranks squared (Σd^2) was calculated to be 1000.

Questions

1. Substitute the correct values into the formula on the right to calculate rho. (*2 marks*)
2. Estimate the value of rho based on the values in the formula for Q1. (*1 mark*)
3. Calculate the actual value of rho based on the values in the formula for Q1. Show your calculations. (*3 marks*)
4. Explain whether or not the calculated value of rho in Q3 is significant. (*2 marks*)
5. What conclusion can be drawn from your answer to Q4? (*2 marks*)

Spearman's rho: A worked example

Why Spearman's rho?

Spearman's is a test of **correlation** between two sets of values. The test is selected when one or both of the variables are **ordinal level** (though it can be used with interval data). The type of design is not an issue here as the investigation is correlational rather than experimental.

In this worked example, we are looking for a **positive correlation** between the attractiveness ratings given to each member of the couples. The level of measurement is ordinal as data is based on scores on an 'unsafe' scale (subjective ratings of attractiveness) which are converted to ranks for the purposes of the test.

The aim...

A study of relationships was conducted to investigate the **matching hypothesis** (Walster *et al.* 1966, see page 122) which proposes that couples in a long-term relationship tend to be similar in terms of physical attractiveness. Twelve couples were selected for the study. Each partner had their photograph taken and these photographs were placed in a random order so it was not obvious who was in a relationship with whom.

The 24 photographs were then given to 20 participants (who had never met any of the couples before). The participants were asked to rate the person in each photograph – out of 20 – in terms of their physical attractiveness. The median attractiveness rating for each photograph was calculated to see if there was a **significant** correlation between pairs in a couple.

The hypotheses...

Alternative hypothesis: *There is a positive correlation between ratings of physical attractiveness given to two partners in a relationship.* **(directional, one-tailed)**

Null hypothesis: *There is no correlation between ratings of physical attractiveness given to two partners in a relationship.*

Step 1: The table of ranks...

Rank each set of scores separately in each group/condition (in this case, for each partner in the couple) from lowest to highest. As before, if two or more scores share the same ranks, find the **mean** of their total ranks.

Step 2: Calculate the difference...

Find the difference between each pair of ranks and square the difference (as shown in the table below). Finally add the differences up, Σ means 'sum of'.

Table 1 Calculations table

Median physical attractiveness rating for female (out of 20)	*Rank for female partner*	*Median physical attractiveness rating for male (out of 20)*	*Rank for male partner*	*Difference between ranks* (d)	d^2
12.5	8	11	2.5	5.5	30.25
16	10	12	4.5	5.5	30.25
13	9	13	6.5	2.5	6.25
8.5	2	14.5	9	–7	49
12	7	15	10.5	–3.5	12.25
10	4.5	7	1	3.5	12.25
11.5	6	13.5	8	–2	4
7	1	15	10.5	–9.5	90.25
9	3	11	2.5	0.5	0.25
17	11	18.5	12	–1	1
18	12	12	4.5	7.5	56.25
10	4.5	13	6.5	–2	4
				Σd^2 =	296

Step 3: Working out the value of rho...

$$\text{rho} = 1 - \frac{6\Sigma d^2}{N(N^2-1)} = 1 - \frac{6 \times 296}{12(144-1)} = 1 - \frac{1776}{1716} = -.035$$

Step 4: The calculated and critical values...

The **calculated value** of rho is –.035

The **critical value** of rho (in Table 2) for a one-tailed test at the 0.05 level where $N = 12$ is .503

As the calculated value of rho (ignoring the sign) is less than the critical value ($p \leq 0.05$) the result is not significant and we must accept the null hypothesis: *There is no correlation between ratings of physical attractiveness given to two partners in a relationship ($p \leq 0.05$).*

In addition the result is actually in the wrong direction (negative rather than positive) and so the hypothesis would not be accepted even if the calculated value was sufficiently large.

Pearson's r: A worked example

Why Pearson's?

Pearson's is a test of correlation between two sets of values. This test is selected when the data are **interval level**. The type of design is not an issue here as the investigation is correlational rather than experimental.

In this worked example, we are looking for a positive correlation between the length of time (in days) spent using biofeedback and the reduction in resting heart rate (measured in beats per minute, bpm). The level of measurement is interval as data is based on 'safe' mathematical (public measurement) scales. The investigation meets the criteria for a **parametric test**.

The aim...

An investigation into **stress** was carried out to see if there is a relationship between the length of time using **biofeedback** (see page 274) and resting heart rate (bpm). Ten participants suffering from chronic stress who had all been using biofeedback for varying lengths of time were selected for the study.

The researchers hypothesised that those who had been using the technique for the longest would have experienced the biggest reduction in their resting heart rate. Medical records were checked so that the participants' baseline heart rate (before using biofeedback) could be compared with their present heart rate to work out the reduction. This reduction was correlated with the length of time (in days) that they had been using biofeedback.

The hypotheses...

Alternative hypothesis: *There is a positive correlation between the number of days participants have been using biofeedback and the reduction in their resting heart rate (bpm).* (**directional, one-tailed**)

Null hypothesis: *There is no correlation between the number of days participants have been using biofeedback and the reduction in their resting heart rate.*

Step 1: The table of data...

In the Table 3 various calculations need to be made for the x and y scores:

- Calculate the sum of the scores for x (Σx) and y (Σy).
- Square each x value and each y value. Calculate Σx^2 and Σy^2.
- Multiply x and y for each participant. Add these values together = $\Sigma(xy)$.

Table 3 Calculations table

Participant	*Days spent using biofeedback (x)*	x^2	*Reduction in heart rate (y)*	y^2	xy
1	4	16	2	4	8
2	7	49	2	4	14
3	15	225	4	16	60
4	22	484	6	36	132
5	23	529	5	25	115
6	32	1024	5	25	160
7	44	1936	2	4	88
8	51	2601	8	64	408
9	62	3844	7	49	434
10	80	6400	8	64	640
	$\Sigma x = 340$	$\Sigma x^2 = 17108$	$\Sigma y = 49$	$\Sigma y^2 = 291$	$\Sigma(xy) = 2059$

Step 2: Working out the value of r...

$$r = \frac{N(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

$$r = \frac{10(2059) - (340)(49)}{\sqrt{(171080 - 115600)(2910 - 2401)}} = \frac{3930}{5314} = .740$$

Step 3: The calculated and critical values...

The **calculated value** of r is .740

The **critical value** of r (in Table 4) for a one-tailed test at the 0.05 level where $df = N - 2 = 8$, is .549.

As the calculated value of r is more than the critical value the result is significant at the 0.05 level and we can reject the null hypothesis and accept the alternative hypothesis: *There is a positive correlation in the number of days participants have been using biofeedback and the reduction in their resting heart rate ($p \leq 0.05$).*

Table 4 Critical values of r

Level of significance for a one-tailed test	0.05	0.025
Level of significance for a two-tailed test	0.10	0.05
df = 2	.9000	.9500
3	.805	.878
4	.729	.811
5	.669	.754
6	.621	.707
7	.582	.666
8	.549	.632
9	.521	.602
10	.497	.576
11	.476	.553
12	.475	.532
13	.441	.514
14	.426	.497
15	.412	.482
16	.400	.468
17	.389	.456
18	.378	.444
19	.369	.433
20	.360	.423
25	.323	.381
30	.296	.349
35	.275	.325
40	.257	.304
45	.243	.288
50	.231	.273
60	.211	.250
70	.195	.232
80	.183	.217
90	.173	.205
100	.164	.195

Calculated value of r must be EQUAL TO or MORE THAN the critical value in this table for significance to be shown.

Apply it

Methods: Using the critical value table

A researcher was interested to know if there was a positive correlation between heat and aggression. The researcher made a note of the average temperature in his local town on various days throughout the year. He also recorded the number of violent incidents that were reported in the local newspapers on those days.

The researcher used a Pearson's test to analyse his data. The calculated value of r was 0.281. Data for daily temperature and number of violent incidents was recorded for 52 days throughout the year.

Questions

1. Is the result significant? Explain your answer. *(3 marks)*
2. What conclusion can be drawn from this study? *(2 marks)*

CHECK IT

1. When would a researcher decide to use a Spearman's rho test? Refer to **two** factors in your answer. *[2 marks]*
2. When would a researcher decide to use a Pearson's r test? Refer to **two** factors in your answer. *[2 marks]*

Test of association: Chi-Squared

Practical activity on page 85

THE SPECIFICATION SAYS

Students should be familiar with the use of inferential tests.

There is one final statistical (inferential) test that you have to study, the Chi-Squared test, which can be used for differences or association.

The key feature of Chi-Squared is that each data item is not listed separately but, instead, a frequency count is given. Usually the data are entered in 4 cells (a 2 x 2 table), but 6 cells or 9 cells etc. can be used and then the contingency table is called 3 x 2 or 3 x 3 respectively. The first number identifies the number of rows and the second number is the number of columns.

The data in each cell must be independent – imagine that each data item is one person, each person can only be placed in one cell of the contingency table (Table 1) below right.

Table 3 Critical values of χ^2

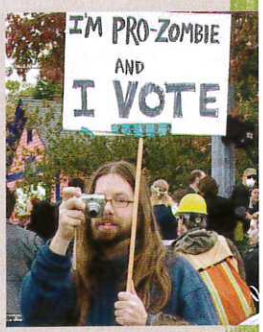
Level of significance for a one-tailed test	0.10	0.05	0.025	0.01
Level of significance for a two-tailed test	0.20	0.10	0.05	0.02
df = 1	1.64	2.71	3.84	5.41
2	3.22	4.60	5.99	7.82
3	4.66	6.25	7.82	9.84
4	5.99	7.78	9.49	11.67

Calculated value of χ^2 must be EQUAL TO or MORE THAN the critical value in this table for significance at the level shown.

Apply it

Methods: Calculating Chi

A researcher wanted to see whether there was an association between age and voting preference in the General Election. One hundred voters were classified as either young (under 25) or old (over 60). Of the 50 'young' voters, 42 voted for the Pro-Zombie Party and 8 for the Anti-Zombie Party. Of the 50 'old' voters, 32 voted for the Anti-Zombie Party and 18 for the Pro-Zombie Party.



Questions

1. Construct a 2 × 2 contingency table for the data above. (*3 marks*)
2. Calculate the value of χ^2 for the data above. (*3 marks*)
3. Explain whether the value of χ^2 you calculated in Q2 is significant. (*2 marks*)
4. Suggest *one* conclusion that could be drawn from your answer to Q3. (*2 marks*)

Chi-Squared

Why Chi?

Chi-Squared is a test of *difference* or *association*. The data are **nominal** and recorded as a frequency count of the categories.

In this worked example, we are looking for a difference in the ability to **decentre** in children aged 5 and children aged 8. There are two **independent groups** of children which means the design is **unrelated**. Finally, the level of measurement is nominal as data is collected in the form of frequencies in two categories: ability to decentre or not.

The aim...

A study of **cognitive development** was conducted to see if there was a difference in children's ability to decentre (see the world from the perspective of another) depending on their age. A group of 5-year-olds and 8-year-olds were given the **three mountains task** (see page 180) to see whether they could choose a card that corresponded to a doll's view rather than their own.

The hypotheses...

Alternative hypothesis: *More 8-year-olds than 5-year-olds are able to select a card that represents a perspective different from their own.* (**directional, one-tailed**)

Null hypothesis: *There is no difference between the number of 5-year-olds and 8-year-olds who can select a card that represents a perspective different from their own.*

Step 1: A 2 × 2 contingency table...

Draw a 2 × 2 contingency table showing the *observed frequencies* (i.e. the data that was collected) in each cell and calculate the totals for each row, each column and the overall total.

Table 1 Contingency table

	5-year-olds	8-year-olds	Totals
Decentre	6 (cell **A**)	28 (cell **B**)	34
Could not decentre	27 (cell **C**)	9 (cell **D**)	36
Totals	33	37	70

Step 2: The table of expected frequencies...

Expected frequencies (E) are now calculated for each of the four cells in the 2 × 2 table. An expected frequency is the frequency that would be expected if there was no difference between the two groups *(if the age of the child had no effect on their ability to decentre)*. The expected frequency is calculated for each cell by multiplying the total for the row by the total for the column divided by the grand total of 70 (taking the data from Table 1).

This calculation is done as shown below, taking the data from Table 1.

O = observed frequencies from the table in Step 1.

Table 2 Calculations table

	E	E–O	$(E-O)^2$	$(E-O)^2 / E$
Cell A	34 × 33 / 70 = 16	6 – 16 = –10	100	6.3
Cell B	34 × 37 / 70 = 18	28 – 18 = 10	100	5.6
Cell C	36 × 33 / 70 = 17	27 – 17 = 10	100	5.9
Cell D	36 × 37 / 70 = 19	9 – 19 = –10	100	5.3

Answers have been rounded to the nearest whole number, except in the final column where they are rounded to one decimal place. This has been done to save space, normally you should work to two or even three decimal places.

Step 3: Working out the value of χ^2...

Add up the values in the final column.

The **calculated value of** χ^2 is 23.1.

Step 4: The calculated and critical values...

To find the **critical value**, calculate the **degrees of freedom** (*df*) by multiplying *(rows – 1)* × *(columns – 1)* = 1. ('Rows' and 'columns' refers to the contingency table.)

The critical value of χ^2 (in Table 3) for a one-tailed test at the 0.05 level, where *df* =1, is 2.71.

As the calculated value of χ^2 is more than the critical value ($p \leq 0.05$) we can reject the null hypothesis and accept the alternative hypothesis: *There is a difference in the number of 5-year-olds and 8-year-olds who can select a card that represents a perspective different from their own* ($p \leq 0.05$).

Reporting psychological investigations

The specification says...

Reporting psychological investigations. Sections of a scientific report: abstract, introduction, method, results, discussion and referencing.

When psychologists come to write up their research for publication in journal articles, they use a conventional format. On this half-spread we describe each of the sections that make up a scientific report.

Key terms

Abstract – The key details of the research report.

Introduction – A look at past research (theory and/or studies) on a similar topic. Includes the aims and hypothesis.

Method – A description of what the researcher(s) did, including design, sample, apparatus/materials, procedure, ethics.

Results – A description of what the researcher(s) found, including descriptive and inferential statistics.

Discussion – A consideration of what the results of a research study tell us in terms of psychological theory.

References – List of sources that are referred to or quoted in the article, e.g. journal articles, books or websites, and their full details.

Study tips

Try it! There is no formal requirement to complete coursework for A level Psychology as there used to be. However, we would definitely recommend that you carry out as many practical investigations as you can. This will give you vital understanding of issues involved in the design of studies, as well as the techniques involved in collecting, summarising and analysing data, and will be of great help to you when it comes to tackling Research Methods questions.

Why not write up one of your investigations in the conventional report format described here? Use one of the practical activities suggested in this book or make up your own (having checked with your teacher that what you propose to do is ethical of course!).

Check it

1. When would a researcher decide to use a Chi-Squared test? Refer to **three** factors in your answer. *[3 marks]*
2. Briefly outline what information psychologists should include within an abstract when reporting psychological investigations. *[3 marks]*
3. Identify and outline **two** sections of a scientific report. *[6 marks]*
4. List **four** sub-sections that should be included within the method section of a psychological report. *[4 marks]*

Sections of a scientific report

Abstract

The first section in a journal article is a short summary/**abstract** (150–200 words in length) that includes all the major elements: the aims and hypotheses, method/procedure, results and conclusions. When researching a particular topic, psychologists will often read lots of abstracts in order to identify those investigations that are worthy of further examination.

Introduction

The **introduction** is a literature review of the general area of investigation detailing relevant theories, concepts and studies that are related to the current study. The research review should follow a logical progression – beginning broadly and gradually becoming more specific until the **aims** and **hypotheses** are presented.

Method

Split into several sub-sections, the **method** should include sufficient detail so that other researchers are able to precisely **replicate** the study if they wish:

- *Design* – The design is clearly stated, e.g. independent groups, naturalistic observation, etc., and reasons/justification given for the choice.
- *Sample* – Information related to the people involved in the study: how many there were, biographical/demographic information (as long as this does not compromise anonymity) and the **sampling method** and **target population**.
- *Apparatus/materials* – Detail of any assessment instruments used and other relevant materials.
- *Procedure* – A 'recipe-style' list of everything that happened in the investigation from beginning to end. This includes a verbatim record of everything that was said to participants: **briefing**, **standardised instructions** and **debriefing**.
- *Ethics* – An explanation of how these were addressed within the study.

Results

The **results** section should summarise the key findings from the investigation. This is likely to feature **descriptive statistics** such as tables, graphs and charts, measures of central tendency and measures of dispersion.

Inferential statistics should include reference to the choice of **statistical test**, **calculated** and **critical values**, the **level of significance** and the final outcome, i.e. which hypothesis was rejected and which retained.

Any **raw data** that was collected and any calculations appear in an appendix rather than the main body of the report.

If the researcher has used **qualitative methods** of research, the results/findings are likely to involve analysis of themes and/or categories.

Discussion

There are several key elements in the **discussion** section. The researcher will summarise the results/findings in verbal, rather than statistical, form. These should be discussed in the context of the evidence presented in the introduction and other research that may be considered relevant.

The researcher should be mindful of the limitations of the present investigation and discuss these as part of this section. This may include reference to aspects of the method, or the sample for instance, and some suggestions of how these limitations might be addressed in a future study.

Finally, the wider implications of the research are considered. This may include real-world applications of what has been discovered and what contribution the investigation has made to the existing knowledge-base within the field.

Referencing

Full details of any source material that the researcher drew upon or cited in the report. **Referencing** may include journal articles, books, websites, etc. Here's an example of a reference from a journal article that appears in the Biopsychology section of this book:

Gupta, S. (1991). Effects of time of day and personality on intelligence test scores. *Personality and Individual Differences, 12(11).* 1227–1231.

Book references take the following format: author(s), date, title of book (in italics), place of publication, publisher. For example:

Flanagan, C. and Berry, D. (2016). *A level Psychology*. Cheltenham: Illuminate Publishing.

Note how the name of the journal and title of the book appear in italics as does the journal volume and issue number (12 and 11 respectively). For a journal article the last information is the page number(s). See more examples of formal academic referencing at the back of this book (page 386).

FEATURES OF SCIENCE

THE SPECIFICATION SAYS...

Features of science: objectivity and the empirical method; replicability and falsifiability; theory construction and hypothesis testing; paradigms and paradigm shifts.

What makes *science* scientific? And is psychology a science? On this spread we attempt to tackle both of these questions by first describing the key features and assumptions of scientific enquiry. We will then consider to what extent psychology as a social scientific discipline rather than a 'natural' science meets these criteria.

KEY TERMS

Paradigm – A set of shared assumptions and agreed methods within a scientific discipline.

Paradigm shift – The result of a scientific revolution: a significant change in the dominant unifying theory within a scientific discipline.

Objectivity – When all sources of personal bias are minimised so as not to distort or influence the research process.

The empirical method – Scientific approaches that are based on the gathering of evidence through direct observation and experience.

Replicability – The extent to which scientific procedures and findings can be repeated by other researchers.

Falsifiability – The principle that a theory cannot be considered scientific unless it admits the possibility of being proved untrue (false).

STUDY TIP

A word about hypotheses.

*We have distinguished between the **null hypothesis** and the **alternative hypothesis** (which might be **one-tailed** or **two-tailed** depending on the aim of the research). An alternative hypothesis might also – alternatively(!) – be referred to as a research **hypothesis**. If a researcher is using an experiment to investigate the hypothesis, the research hypothesis may be referred to as an **experimental hypothesis**. Or if the method of research is a correlation, the research hypothesis is a **correlational hypothesis**. Phew! That was probably a bit more than a word...*

Features of science

Paradigms and paradigm shifts

The philosopher Thomas Kuhn (1962) suggested that what distinguishes scientific disciplines from non-scientific disciplines is a shared set of assumptions and methods – a **paradigm**. Kuhn suggested that social sciences (including psychology) lack a universally accepted paradigm and are probably best seen as 'pre-science' as distinct from natural sciences such as biology or physics. Natural sciences are characterised by having a number of principles at their core such as the theory of evolution in biology, or the standard model of the universe in physics. Psychology, on the other hand, is marked by too much internal disagreement and has too many conflicting approaches to qualify as a science and therefore is a pre-science (this view of psychology has been challenged – see below).

According to Kuhn, progress within an established science occurs when there is a scientific revolution. A handful of researchers begin to question the accepted paradigm, this critique begins to gather popularity and pace, and eventually a **paradigm shift** occurs when there is too much contradictory evidence to ignore. Kuhn cited the change from a Newtonian paradigm in physics towards Einstein's theory of relativity as an example of a paradigm shift.

Theory construction and hypothesis testing

Science tests theories – but what is a **theory**? A theory is a set of general laws or principles that have the ability to explain particular events or behaviours. *Theory construction* occurs through gathering evidence via direct observation (see the **empirical method** on the facing page). For instance, I may have a 'hunch' that **short-term memory** has a limited capacity based on the observation that people struggle to remember much when they are 'bombarded' with information. A series of **experiments** reveals that the average short-term memory span is around 7 (give or take 2) items of information. Let's call this *Berry's Law* ... OK fine, someone else got there first – but this is a good example of a theory as it proposes a simple and economical principle which appears to reflect reality. It provides understanding by explaining regularities in behaviour.

It should also be possible to make clear and precise predictions on the basis of the theory. This is the role of hypothesis testing. An essential component of a theory is that it can be scientifically tested. Theories should suggest a number of possible hypotheses – for instance, *Berry's Law* (see – it's catching on ...) suggests that people will remember 7-digit postcodes more effectively than 14-digit mobile phone numbers. A hypothesis like this can then be tested using systematic and objective methods to determine whether it will be supported or refuted. In the case of the former, the theory will be strengthened; in the case of the latter, the theory may need to be revised or revisited. The process of deriving new hypotheses from an existing theory is known as deduction.

Apply it — Concepts: Does psychology have a paradigm?

Kuhn's argument was that psychology's lack of an accepted paradigm means it is yet to achieve the status of normal science, and is instead, pre-science. Certainly there are a number of theoretical perspectives in psychology that have suggested quite different ideas and ways of investigating the human subject.

However, not all commentators agree with Kuhn's conception of psychology as pre-scientific. For instance, the vast majority of researchers would accept a definition of psychology as *the study of mind and behaviour* suggesting there is broad agreement. Similarly, it could be argued that psychology has already progressed through several paradigm shifts from Wundt's early **structuralism** to the dominant **cognitive neuroscience** model of today.

Finally, several researchers (including Feyerabend 1975) have suggested that Kuhn's conception of 'proper' science as orderly and paradigmatic is flawed – and that most sciences are in fact characterised by internal conflict, dispute and a refusal to accept new ideas in the face of evidence.

Questions

1. Choose *two* approaches in psychology and explain how the main assumptions and methods of enquiry within these two approaches differ.
2. Use your knowledge of the historical development of psychology to explain how the discipline may have experienced several paradigm shifts.

Falsifiability

Another philosopher of science whose work appeared around the same time as Thomas Kuhn was Karl Popper (1934) who argued that the key criterion of a scientific theory is its **falsifiability**. Genuine scientific theories, Popper suggested, should hold themselves up for **hypothesis** testing and the possibility of being proven *false*. He believed that even when a scientific principle had been successfully and repeatedly tested, it was not necessarily true. Instead it had simply not been proven false – yet! This became known as the *theory of falsification*. Popper drew a clear line between good science, in which theories are constantly challenged, and what he called 'pseudosciences' which couldn't be falsified.

Those theories that survive most attempts to falsify them become the strongest – not because they are necessarily true – but because, despite the best efforts of researchers, they have not been proved false. This is why psychologists avoid using phrases such as '*this proves*' in favour of '*this supports*' or '*this seems to suggest*' – and why, as we have seen, an alternative hypothesis must always be accompanied by a **null hypothesis**.

Replicability

An important element of Popper's **hypothetico-deductive method** (described above) is **replicability**. If a scientific theory is to be 'trusted', the findings from it must be shown to be repeatable across a number of different contexts and circumstances.

Replication has an important role in determining the **validity** of a finding. We have already discussed the role of **replication** in determining the **reliability** of the *method* used in a study (see pages 66–69). Replication is also used to assess the validity of a *finding*; by repeating a study, as Popper suggests, over a number of *different* contexts and circumstances then we can see the extent to which the findings can be **generalised**. In order for replicability to become possible, it is vital that psychologists report their investigations with as much precision and rigour as possible, so other researchers can seek to *verify* their work and verify the findings they have established.

Objectivity and the empirical method

Scientific researchers must strive to maintain **objectivity** as part of their investigations. In other words, they must keep a 'critical distance' during research. They must not allow their personal opinions or biases to 'discolour' the data they collect or influence the behaviour of the participants they are studying. As a general rule, those methods in psychology that are associated with the greatest level of **control** – such as **lab experiments** – tend to be the most objective.

Objectivity is the basis of the **empirical method**. The word *empiricism* is derived from the Greek for 'experience' and empirical methods emphasise the importance of data collection based on direct, sensory experience. The **experimental method** and the **observational method** are good examples of the empirical method in psychology. Early empiricists such as John Locke saw knowledge as determined only by experience and sensory perception. Thus, a theory cannot claim to be scientific unless it has been empirically tested and verified.

'It is a capital mistake to theorise before you have all the evidence. It biases the judgment.'

Sherlock Holmes, *A Study in Scarlet*

Apply it

Concepts: Psychology as a science: the case for ...

Scientific psychology lifts everyday understanding of human behaviour above the level of commonsense. Critics of psychology may claim it amounts to little more than commonsense, but many key findings in psychology are counter-intuitive and not what a commonsense view would predict.

By adopting a scientific model of enquiry, psychology gives itself greater credibility by being placed on equal footing with other, more established sciences (despite Kuhn's suggestion that psychology is just a pre-science).

The scientific approach in psychology has provided many practical applications that have improved people's lives and challenged/modified dysfunctional behaviour.

Questions

1. As an example of counter-intuitive findings, explain why Milgram's findings were not what most people would have predicted.
2. List at least *two* of the practical applications of psychology and examine their effectiveness.

Apply it

Concepts: Psychology as a science: the case against ...

Although many psychologists try to maintain objectivity within their research, some of the methods that psychologists use are **subjective**, non-standardised and unscientific.

Science is based on the assumption that it is possible to produce universal laws that can be generalised across time and space. However, this may not be possible in psychology: samples of participants in studies are rarely representative and conclusions drawn may often be influenced by cultural and social norms.

Much of the subject matter in psychology cannot be directly observed and must be based on **inference** rather than objective measurement.

Questions

1. Provide an example of subjective methods in psychology, with reference to specific studies.
2. Even when more objective methods are used, explain why objectivity may be much harder to achieve in psychology than in other sciences, e.g. physics, chemistry.
3. Why might replicability be harder to achieve in psychology than other sciences?
4. Explain which psychological approaches this most applies to and why.
5. Explain why many findings gained from experimental research may lack **ecological validity** and/or **temporal validity**. Give some examples.
6. Explain why the issue of inference is a criticism that may be levelled at the **cognitive approach**.

CHECK IT

1. Outline what is meant by *replicability* and *falsifiability* in psychological research. *[4 marks]*
2. Outline what is meant by the following terms in scientific research: (i) paradigm (ii) paradigm shift. *[4 marks]*
3. Briefly discuss the importance of theory construction and hypothesis testing in scientific research. *[6 marks]*
4. Briefly discuss arguments for and against the idea that psychology is a science. *[10 marks]*

Revision summaries

Correlations

Revisiting the analysis of co-variables.

Analysis and interpretation of correlations

Correlations and correlation coefficients
Relationship between two continuous co-variables.
Correlation coefficient represents strength and direction of relationship.

Working out what a coefficient means
The closer the coefficient is to −1 or +1, the stronger the relationship.

Case studies and content analysis

Two forms of research method.

Case studies

Case studies
Detailed analysis of an unusual individual or event, e.g. the London riots.

Characteristics
Tend to produce qualitative data, and be longitudinal.

Evaluation

Strengths
Insight into unusual cases, e.g. HM may provide understanding of normal functioning.
Generate hypotheses for future studies.

Limitations
Generalisation is a problem and conclusions based on subjective interpretation of the researcher.

Content analysis

Content analysis
A form of observation in which communication is studied indirectly.

Coding and quantitative data
Data must be categorised into meaningful units (and then analysed by counting words, etc).

Thematic analysis and qualitative data
Recurrent ideas that keep 'cropping up' in the communication are described.

Evaluation

Strengths
Fewer ethical issues and high external validity.

Limitations
Information may be studied out of context and descriptive forms of analysis may be less objective.

Reliability

A measure of consistency.

Reliability

Introducing reliability
Psychologists tend not to measure concrete things so reliability is difficult to establish.

Test–retest
The same test is administered to the same person (or group) on different occasions and results compared.

Inter-observer reliability
Observers should compare data in a pilot study or at end of actual study to make sure behavioural categories are consistently applied.

Improving reliability

Questionnaires
If a questionnaire has low test–retest reliability, some items may need to be changed to closed questions as these may be less ambiguous.

Interviews
Should avoid questions that are leading or ambiguous and ensure interviewers are trained.

Experiments
Standardisation of procedures will minimise extraneous variables.

Observations
Behavioural categories should be properly operationalised and not overlap.

Types of validity

A measure of 'truth'.

Validity

Introducing validity
Whether a test, scale, etc, produces a legitimate result which represents behaviour in the real world.

Internal and external validity
Whether something measures what it was designed to measure, and whether findings can be generalised.

Ecological validity
The extent to which findings can be generalised from one setting to other settings.
Mundane realism of task may affect ecological validity.

Temporal validity
Do findings from a study hold true over time?

Assessment of validity

Face and concurrent validity
Does a test measure what it is supposed to 'on the face of it'?
Do results match with a previously established test?

Improving validity

Experimental research
Use of a control group.
Standardised procedures.
Single-blind and double-blind trials.

Questionnaires
Use of lie scales and anonymity to reduce social desirability.

Observations
Findings may be more authentic in covert observations.

Qualitative methods
Depth and detail may increase validity but further enhanced through triangulation.

Choosing a statistical test

Statistical tests tell us whether results are significant.

Choosing a statistical test

Statistical testing
Determine whether we can accept or reject the null hypothesis.

Difference or correlation
Correlation includes tests of association (Chi-Squared).

Experimental design
Related (repeated measures or matched pairs) or unrelated (independent groups).

Parametric tests
Interval level data, normal distribution and homogeneity of variance.

Levels of measurement

Nominal data
Data represented in the form of categories, e.g. counting how many boys and girls in a year group – boys and girls are discrete categories.

Ordinal data
'Unsafe' data which can be placed in rank order, e.g. rating your liking of psychology on a scale of 1–10.

Interval data
Based on numerical and public scales of measurement with units of equal size, e.g. length, temperature.

Probability and significance

Psychological research works on probabilities rather than certainties.

Probability and significance

The null hypothesis
The null hypothesis states no difference between conditions. Statistical tests determine whether this should be accepted or rejected.

Levels of significance and probability
The significance level is the point at which the researcher can accept the alternative hypothesis (usually 5% in psychology).

Use of statistical tables

Calculated and critical values
The calculated value must be compared with a critical value to determine significance.

Using tables of critical values
Is the test one-tailed or two-tailed?
What is the N value?
Which level of significance?

Lower levels of significance
A more stringent level, e.g. 1%, should be used when research has a human cost or the study is a one-off.

Type I and Type II errors

Type I error
The incorrect rejection of a true null hypothesis.

Type II error
The incorrect acceptance of a false null hypothesis.

Different Statistical tests

Formula for determining significance.

Tests of difference

Mann–Whitney
Test of difference between two sets of data.
Unrelated design.
Data at least ordinal level.

Wilcoxon
Test of difference between two sets of data.
Related design.
Data at least ordinal level.

Parametric tests of difference

Unrelated *t*-test
Test of difference between two sets of data.
Unrelated design.
Data at interval level.
Data drawn from normally distributed sample population and homogeneity of variance.
Homogeneity of variance.

Related *t*-test
Test of difference between two sets of data.
Related design.
Data at interval level.
Data drawn from normally distributed sample population and homogeneity of variance.
Homogeneity of variance.

Tests of correlation

Spearman's
Test of correlation between co-variables.
Data at least ordinal level.

Pearson's
Test of correlation between co-variables.
Data at interval level.
Data drawn from a normally distributed population and homogeneity of variance.

Test of difference/association

Chi-Squared
Test of difference between two sets of data or association between co-variables.
Data is independent.
Nominal data.

Rule of R

Rule of R
Tests with a letter 'R' in their name are those where the calculated value must be equal to or more than the critical value.

Features of science

What makes science scientific?

Paradigms and paradigm shifts
Scientific subjects have a shared set of assumptions and a scientific revolution occurs when there is a paradigm shift.

Theory construction and hypothesis testing
Theory construction occurs through gathering evidence from direct observation.
Researchers can produce clear and precise hypotheses to test the validity of the theory.

Falsifiability
Scientific theories should hold themselves up for hypothesis testing and the possibility of being proved false.

Replicability
If a scientific theory is to be 'trusted' (i.e. valid), its findings must be shown to be repeatable across time and context.
The methods used should also be repeatable, i.e. reliable.

Objectivity and the empirical method
Scientists must minimise all sources of personal bias and gather evidence through direct observation and experience.

Reporting psychological investigations

Psychologists use a conventional format when presenting their research.

Abstract
A short summary of the different elements in the report.

Introduction
Literature review including aim and hypothesis.

Method
Includes design, sample, apparatus/materials, procedure, ethics.

Results
Descriptive and inferential statistics.

Discussion
Analysis of results, limitations and wider implications.

References
List of sources (journal articles, books, web sources).