A LEVEL RESEARCH METHODS REVISION
CASE STUDIES

- These are detailed studies of a single individual, institution or event
- It is an example of evidence-based research (psychologists turn to individual cases partly to look at unusual behaviours, and partly to look in greater detail at any kind of behaviour)
- The case study is a scientific research method and so aims to use objective and systematic methods
- Case studies use information from a range of sources, like from the person concerned and also from their family and friends
CASE STUDIES

• Many techniques may be used (the people may be interviewed or observed in daily life)

• Psychologists might use IQ or personality tests or a questionnaire to produce psychological data about the target person or group of people

• They may use the experimental method to test what the target person/group can or cannot do

• The findings are organised into themes to represent the individual’s thoughts, emotions, experiences and abilities

• The data therefore may be presented in a qualitative way, though quantitative data may also be included (like scores from psychological tests)

• Case studies are generally longitudinal (i.e. they follow the individual or group over an extended period of time)
CASE STUDIES

• Examples of case studies are:
  – Henry Molaison (HM) – who had his hippocampus removed to reduce epileptic fits and resulted in an inability to form new long-term memories
  – Little Hans – to illustrate the principle of psychoanalysis (Little Hans demonstrated the oedipus complex)
  – Phineas Gage – an explosion of dynamite drove a tamping iron through his skull. He survived despite losing a great deal of brain matter and was able to function fairly normally. This was important in the development of brain surgery to remove tumours as it showed that parts of the brain could be removed without having a fatal effect
CASE STUDIES - EVALUATION

STRENGTHS

• This method offers rich, in-depth data
• This data can provide insights into the complex interactions of many factors, in contrast with experiments where many variables are held constant
• This means that insights overlooked using other methods are likely to be identified
• Case studies can be used to investigate human behaviour and experience that are rare
• E.g. investigating cases of children locked in a room during their childhood (Genie)
• Such cases enable researchers to see what effects such as disruption of attachment have on emotional development
• It would not be ethical to generate such conditions experimentally
CASE STUDIES - EVALUATION

LIMITATIONS

• It is difficult to generalise from individual cases as each one has unique characteristics

• E.g. HM told us a great deal about the effects of his operation on memory, but we do not know to what extent his epilepsy rather than the brain damage may have affected aspects of his behaviour

• There are important ethical issues such as confidentiality and informed consent

• Many cases are easily identifiable due to their unique characteristics, even when real names are not given

• Many individuals like HM or Little Hans are not able or not asked to give informed consent

• Psychological harm may also be an issue when an individual like HM is tested repeatedly over decades

• It is often necessary to use recollection of past events as part of the case history, which may be unreliable

• Researchers may lack objectivity as they get to know the case
A LEVEL RESEARCH METHODS

CONTENT ANALYSIS
CONTENT ANALYSIS

• Conducting this is similar to any observational study, but instead of observing actual people a researcher usually makes observations indirectly through books, films, advertisements and photographs (an artefact people have produced)

• As in an observational study, the researcher has to make design decisions about the sampling method, coding the data and method of representing data
CONTENT ANALYSIS

SAMPLING METHOD

• The researcher has to decide the following:
  – If analysing the content of books, does the researcher look at every page or just every nth page (like time sampling)?
  – If comparing the content in various books, does the researcher select books randomly from a library or identify certain characteristics (e.g. look at books that are biographies or romantic fiction)?
  – If analysing adverts on TV, does the researcher sample behaviours every 30 seconds for example, or note whenever certain behaviours occur?
CONTENT ANALYSIS

METHOD OF REPRESENTING DATA

• Data can be recorded in each behavioural category in two different ways *(qualitatively and quantitatively)*

• If we were to perform a content analysis of paintings then we would first identify the behavioural categories and then record instances in each category:
  – We can count instances = a quantitative analysis
  – We can describe examples in each category = a qualitative analysis
CONTENT ANALYSIS
CODING THE DATA

• The researcher uses behavioural categories
• For example, if the researcher wishes to look at the way men and women are portrayed in books, they create a list of behavioural categories and then count instances
• Decisions about behavioural categories may involve a **thematic analysis**
CONTENT ANALYSIS - EVALUATION

STRENGTHS

• Content analysis tends to have high ecological validity as it is based on observations of what people actually do (real communications that are current and relevant, like books that people read)

• When sources can be retained or accessed by others, the content analysis can be replicated, and therefore the observations can be tested for reliability

LIMITATIONS

• Observer bias reduces the objectivity and validity of findings as different observers may interpret the meaning of the behavioural categories differently

• Content analysis is likely to be culture biased as interpretation of verbal or written content will be affected by the language and culture of the observer and the behavioural categories used
THEMATIC ANALYSIS

• One problem with qualitative data is that it is difficult to summarise
• It cannot be analysed like quantitative data can (e.g. with measures of central tendency and dispersion, and also with graphs)
• Instead, qualitative data is summarised by identifying themes in the material to be analysed
• The material to be analysed might be a book, adverts on TV, a transcript from an interview, or a researcher might want to analyse graffiti or analyse videotaped play sessions with children
THEMATIC ANALYSIS

• Thematic analysis is a very lengthy process as it is thorough and repetitive (every item is carefully considered and the data are gone through repeatedly)

• The main intentions are:
  – To impose some kind of order on the data
  – To ensure that the “order” represents the participants’ perspective, i.e. fits in with how they see the phenomenon
  – To ensure that this “order” emerges from the data rather than any preconceptions
  – To summarise the data so that hundreds of pages of text or hours of videotapes can be reduced
  – To enable themes to be identified and general conclusions drawn

• There is no one method to use, but the table on the next slide gives a general picture of what is done
## THEMATIC ANALYSIS

### GENERAL PRINCIPLES

<table>
<thead>
<tr>
<th>Description</th>
<th>Applied to the Analysis of Graffiti</th>
<th>Applied to the Analysis of Videotaped Play Sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read and re-read the data transcript objectively, trying to understand the meaning communicated and the perspective of the participants. No notes should be made.</td>
<td>Study a photograph/written record of a wide range of graffiti</td>
<td>Watch videos of play sessions/read transcription of what happened (including facial expressions and body movements)</td>
</tr>
<tr>
<td>Break the data into meaningful units.</td>
<td>In the case of graffiti it would be each item of graffiti.</td>
<td>Each verbal and non-verbal movement would constitute a unit.</td>
</tr>
<tr>
<td>Assign a label or code to each unit (these are the initial categories used). Each unit may be given more than one code/label.</td>
<td>Each unit of graffiti is given a code to describe its meaning (e.g. humour, love, power, domination).</td>
<td>Each unit is coded, e.g. “playing with toy”, “sadness expressed” or “request made”.</td>
</tr>
<tr>
<td>Combine simple codes into larger categories/themes and then instances can be counted or examples provided.</td>
<td>Larger categories are developed which combine units, like “interpersonal concerns”.</td>
<td>Larger categories developed such as “negative emotion”.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A check can be made on the emergent categories by collecting a new set of data and applying categories. They should fit the new data well if they represent the topic area investigated.</td>
</tr>
</tbody>
</table>
A LEVEL RESEARCH METHODS

RELIABILITY
RELIABILITY OF OBSERVATIONAL TECHNIQUES

• Observations are a form of measurement where researchers record what people/animals are doing

• The researcher will keep a record of events observed using a set of behavioural categories

• An example is below of playground behaviours observed at break time:

<table>
<thead>
<tr>
<th>BEHAVIOUR</th>
<th>BOYS</th>
<th>GIRLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Running</td>
<td>⌂</td>
<td>⌂</td>
</tr>
<tr>
<td>Skipping</td>
<td>I</td>
<td>⌂</td>
</tr>
<tr>
<td>Talking</td>
<td>⌂</td>
<td>⌂</td>
</tr>
<tr>
<td>Laughing</td>
<td>⌂</td>
<td>⌂</td>
</tr>
<tr>
<td>Shouting</td>
<td>⌂</td>
<td>I</td>
</tr>
</tbody>
</table>
RELIABILITY OF OBSERVATIONAL TECHNIQUES

Assessing Reliability

• It is important that the record is a reliable measurement of behaviour
  – In order to do this the observer repeats the observations a second time (e.g. by watching a video recording)
• If the observations are reliable, then the second set of observations should be more or less the same as the first
• The issue with this is that the observer may be biased and so what they think they see may be affected by their beliefs/stereotypes
• A better way to assess accuracy is to have two or more observers making separate recordings and then compare these records
• The extent to which the observers agree on the observations they record is called **inter-observer reliability**
  – This can be calculated as a correlation coefficient for pairs of scores
  – A result of 0.80 or more suggests good inter-observer reliability
RELIABILITY OF OBSERVATIONAL TECHNIQUES

Improving Reliability: Behavioural Categories

• If the score for inter-observer reliability is low, there are ways to improve this
  – It may be that the behavioural categories were not operationalised clearly enough, so one observer may have interpreted an action as one thing (e.g. “talking”) but another may have interpreted it as another thing (e.g. shouting”). So behaviour categories need to be clearer
  – It may be that some observers just need more practice using the behavioural categories so they can respond more quickly
RELIABILITY OF SELF-REPORT TECHNIQUES

• These research methods are where an individual tells the researcher what they think or feel (e.g. questionnaire or interview)

• Psychology tests like IQ or personality tests are similar to self-report techniques and their reliability is particularly important

• We are measuring an aspect of a person and want to ensure that the measurement is reliable (i.e. we would get the same set of answers or the same score each time the test was taken)

• In some cases the score might vary
  – E.g. taking a test that measures mood will change from week to week and so we need to be sure that any change in the score is due to the person and not the test
RELIABILITY OF SELF-REPORT TECHNIQUES

Assessing Reliability: Test-retest Reliability

• This is used to assess the reliability of a psychological test or other self-report measures.

• The test/questionnaire designer gives the test to a group of people and then gives the same people the same test a second time.

• Usually there is a short interval between the tests (e.g. a week or two) so that people don’t remember their answers.

• If the measure is reliable the outcome should be the same every time.

• The scores for each person are compared using correlation (like with observational techniques).
RELIABILITY OF SELF-REPORT TECHNIQUES

Assessing Reliability: Inter-interviewer Reliability

• In the case of interviews, a researcher could assess the reliability of one interviewer by comparing answers on one occasion with answers from the same person with the same interviewer a week later.

• Or, the researcher might want to assess the reliability (consistency) of two interviewers using the same method as with two observers.
RELIABILITY OF SELF-REPORT TECHNIQUES

Improving Reliability: Reduce Ambiguity

• Low reliability in a psychology test may be because some test items are ambiguous so people give different answers
  – E.g. a question might be “what are your thoughts about dieting?” Some people might interpret this as being a question asking for factual information, and provide facts about dieting, whereas others might think the question was about emotions and respond with their own feelings about their efforts of dieting

• In such cases test items (or questions on a questionnaire) need to be re-examined and rewritten
RELIABILITY OF EXPERIMENTS

• The dependent variable in an experiment is often measures using a rating scale or behavioural categories, for example:
  – The DV in Bandura’s study using the Bobo doll was the aggressive behaviour of the children. This was assessed by observing their behaviour in a room full of toys and using behavioural categories like verbal imitation
  – One of the DV’s in Rutter and Sonuga-Barke’s study on Romanian orphans was IQ scores

• So, reliability in an experiment may be concerned with whether the method used to measure the DV is consistent; i.e. the observations or the self-report method (the same is true of how co-variables in correlations are measured)
RELIABILITY OF EXPERIMENTS

Improving Reliability: Standardisation

• Another aspect of reliability that is an issue in experiments (and any research study) is the procedures.
• The procedures are often repeated for different participants.
• It is important that the procedures are exactly the same each time (i.e. reliable) because otherwise we can’t compare the performance of participants.
  – For this reason procedures are standardised.
• If another researcher wishes to repeat the experiment, they also need to use exactly the same procedures.
A LEVEL RESEARCH METHODS

VALIDITY
INTERNAL VS EXTERNAL VALIDITY

Validity can be separated into concerns about what goes on within a study (internal validity) and concerns about what goes on after the study (external validity).

Examples of internal validity include:

- **Investigator Effects** – anything that an investigator does that has an effect on a participant’s performance other than what was intended (e.g. encouraging participants to try harder)

- **Demand Characteristics** – cues that inadvertently communicate the aims of the study to participants, such as the Bobo doll in Bandura’s study “inviting” an aggressive response

- **Confounding Variables** – a variable in an experiment that varies systematically with the IV, and therefore conclusions cannot be drawn about what caused changes in the DV

- **Social Desirability Bias** – in a questionnaire, the tendency for participants to provide answers that do not reflect reality because people prefer to show themselves in a good light and don’t always answer questions honestly

- **Poorly Operationalised Behavioural Categories** – observers can’t record reality because the categories are not clear
INTERNAL VS EXTERNAL VALIDITY

• External validity concerns generalising the findings of a study to other people (population validity), historical periods (historical or temporal validity) and settings (ecological validity)

• Ecological validity may be the most important of these and the one that students often misunderstand
ECOLOGICAL VALIDITY

• This refers to being able to generalise the findings from one study to other situations, most particularly to be able to assume that the findings applies to “everyday life”

• The term natural in “natural experiment” leads people to think that such studies must be high in ecological validity

• The fact that a lab is an artificial manufactured environment leads people to assume that these are low in ecological validity

• The issue of “naturalness” is more subtle than that
ECOLOGICAL VALIDITY

Example of a Natural Experiment

• In studies like the ones on the Romanian orphans (institutionalisation), the IV is whether the children were adopted before or after the age of 6 months.
• This variable was not controlled by the researchers, and is sometimes described as varying “naturally.”
• However, this “naturalness” does not make the study high in ecological validity.

The key feature to consider is the DV, not the IV

• In the Romanian orphan study one DV was intellectual development (assessed using IQ tests).
• These tests are conducted in a controlled environment and may be quite “artificial” measures of intellectual development.
• This is what we should consider when assessing the ecological validity of the findings from these studies.

We need to consider the method used to assess the DV rather than the “naturalness” of the IV

• The IV is actually irrelevant to the ecological validity of the findings.
ECOLOGICAL VALIDITY

Example of a Field Experiment

• These are conducted in a natural environment, but that is not what matters when considering ecological validity

• The DV may be important in terms of the task that was used to measure the DV

• For example, Godden and Baddeley’s study on context-dependent forgetting (deep-sea divers), the divers learned a set of words either on land or underwater and then had to recall the word list either on land or underwater

• Consider the questions on the next slide:
ECOLOGICAL VALIDITY

Example of a Field Experiment

1. What environment was the study conducted in?
   – It was either on land or underwater
   – The divers may have felt relaxed underwater and acted as “normal” as they were divers and used to it

2. How was the DV measured?
   – It was measured by learning word lists
   – This is an artificial way to test memory
   – It has low mundane realism, i.e. low similarity to everyday life

3. Were the participants aware their behaviour was being studied?
   – Yes, this means they may not behave “naturally”
   – They may want “to look good” (social desirability bias) or they may try to behave in line with the researcher’s expectations
ECOLOGICAL VALIDITY

Another Field Experiment

• Godden and Baddeley (1980) completed a further study without using word lists; instead participants were trained to do a task that is commonly required of deep-sea divers (they had to learn how to transfer nuts and bolts from one brass plate to another)

• We might consider that this study had higher ecological validity than their first study because the task had greater mundane realism

• However, the participants were still aware of being studied and this may have affected their behaviour
ECOLOGICAL VALIDITY

So........what is Ecological Validity

• It is not easy to decide whether a study has high or low ecological validity

• Think about different aspects of the study, all of which contribute to the question of whether an observed effect can be generalised to other settings, especially “everyday life”

• It is much less about the environment where people are studied and more about
  – How the DV was measured
  – Whether participants knew their behaviour was being assessed
ASSESSING VALIDITY

• If a researcher decides to test whether men or women are more stressed at work and decides to measure stress using a questionnaire

• We can assess the reliability of the measurements and this may be reliable, but it may not be valid (i.e. does it actually measure stress)

• There are a number of ways to assess the validity of this stress questionnaire

• Two methods are:

  1. **Face Validity** – which concerns the issue whether a self-report measure *looks* like it measures what it is intended to measure (e.g. whether the questions on a stress questionnaire are related to stress)

  2. **Concurrent Validity** – which involves comparing the current method of measuring stress with a previously validated one on the same topic. To do this, participants are given both measures at the same time and then their scores are compared. We would expect people to get similar scores on both measurements, thereby confirming concurrent validity of the current questionnaire
IMPROVING VALIDITY

• If the questionnaire is judged to have poor face validity then the questions should be revised so they relate more obviously to the topic

• If concurrent validity is low then the researcher should remove questions which may seem irrelevant and try checking the concurrent validity again

• In the case of internal and external validity issues described previously, improvements should come from better research design
  – For example, double blind can be used to prevent participants guessing research aims (neither the person running the study nor the participant knows the aims of the study)
A LEVEL RESEARCH METHODS

FEATURES OF SCIENCE
WHAT IS A SCIENCE?

• According to the British Science Council the definition of science is:
  – “the pursuit of knowledge and understanding on the natural and social world following a systematic methodology based on evidence”

• Science seeks to explain natural events and make useful predictions.

• Observations, questionnaires, interviews, case studies (i.e. experiments) are all scientific methods
  – They are systematic methods where researchers are concerned about reliability, validity and control
FEATURES OF SCIENCE

• Scientific Knowledge is based on 5 key features:

  **Empirical Evidence**
  • This is the information gained through direct observation or experiment
    – People can claim anything, but the only way we know something is true is through direct testing (**empirical evidence**)

  **Objectivity**
  • Scientists strive to be objective in their work and should not let their own assumptions or expectations affect their records
    – This is why psychologists carefully control conditions (particularly in lab studies) which allows cause and effect relationships to be investigated

  **Replication**
  • The knowledge gained through the scientific process is validated through replication (repeating a study)
    – If the outcome in further studies is the same as the original results then this affirms the original results (especially if they are gained by a different psychologist)
FEATUES OF SCIENCE

Theory Construction
• Facts alone are useless. Explanations or theories need to be constructed to make sense of the facts
  – A Theory is a collection of general principles that explain observations and facts
  – Scientists use both **inductive** and **deductive** methods, so sometimes theory comes before hypothesis testing and sometimes it comes after

Hypothesis Testing
• Theories are modified through the process of hypothesis testing
  – A good theory must be able to generate testable expectations (stated in the form of hypotheses)
  – If a scientist fails to find support for the hypothesis, then the theory needs modification
THE SCIENTIFIC METHOD

• Science is a process which we can obtain information that is valid
• The scientific method starts with observations of a natural phenomenon
• In the **inductive model** this leads to scientists to develop hypotheses which are tested empirically which may lead to new questions and new hypotheses
• Eventually the data may be used to construct a theory
• The **deductive model** places theory construction at the beginning, after making observations
• In both cases the process is repeated over and over again to refine knowledge
• See the diagrams on the next slide
THE SCIENTIFIC METHOD

**Induction**
- Observations
- Testable Hypothesis
- Conduct a study to test the Hypothesis
- Draw Conclusions
- Propose Theory

**Deduction**
- Observations
- Propose Theory
- Testable Hypothesis
- Conduct a study to test the Hypothesis
- Draw Conclusions
FALSIFIABILITY

• Science is a process that is constantly evolving as people realise that it can be done better
• Up until the 1930s scientists believed that their task was to find examples that would confirm their theories
• Popper, a philosopher of science, brought about a major change in the way scientists thought about proof
• He argued that it was not possible to confirm a theory; it was only possible to disconfirm it
• In 1934 he stated that “no matter how many instances of white swans we may have observed, this does not justify the conclusion that all swans are white”
• No number of sightings of white swans can prove the theory that all swans are white, whereas just one sighting of just one black swan will disprove it
• This led to the realisation that the only way to prove a theory correct was actually to seek disproof (falsification), i.e. look for black swans
FALSIFIABILITY

• Therefore, we start research with a null hypothesis – “Not all swans in the world are white”, i.e. there are black swans

• Then we search for swans and record many sightings but if we see no black swans this leads us to be reasonable certain (as we can never be absolutely certain) that the null hypothesis is false

• We can therefore reject the null hypothesis (with reasonable certainty), which means that the alternative hypothesis must be true (alternative to the null)
  – The alternative hypothesis would be “all swans are white”
FALSIFIABILITY

• **Falsifiability** refers to being able to prove a hypothesis wrong

• In any study it is necessary to be able to make a statement (a hypothesis) that can be proved wrong (i.e. a hypothesis that is falsifiable)

• One of the criticisms of some scientific approaches, such as Freudian psychoanalysis, is that they lack falsifiability
Kuhn (1962) argued against the process suggested by Popper’s theory of falsification (whereby theories are fine tuned by a successive series of experiments) but that the world develops through revolutions.

He proposed that there are two main phases in science:

- One is called “normal science” where one theory remains dominant despite occasional challenges from disconfirming studies. Gradually the disconfirming evidence accumulates until the theory can no longer be maintained and then it is overthrown.

- This is the second phase (a revolutionary shift). Kuhn spoke of a **paradigm** rather than a “theory”. He said that a science has a unified set of assumptions and methods.

So Kuhn described science as a set of ideas which remain constant for a period of time then undergoing a “paradigm shift”.

This is when the current theories don’t explain some phenomenon and someone proposes a new theory.

Kuhn found that all scientific fields go through these paradigm shifts many times as new theories supplant the old.
A classic example of a paradigm shift was the revolution in our understanding of the universe due to the work of the Polish astronomer Copernicus in the 16th Century. He overthrew the belief held for nearly 2,000 years that the earth was the centre of the universe. Such changes are not as logical as Popper’s view of science might suggest. According to Kuhn, scientific progress is more like a religious conversation and is related to social factors (e.g. what other people are saying). Kuhn’s view itself is potentially an example of a paradigm shift in the sciences – a change of view from seeing science as logical to science as a social construction (i.e. built through dialogues with other people) – though this view is still hotly debated (Jones and Elcock, 2001).
A LEVEL RESEARCH METHODS

PROBABILITY
PROBABILITY

• If we want to know whether men or women are better map readers, what we actually do is test whether men and women are the **same** in terms of their map reading abilities and seek to falsify the **null hypothesis**
  – A null hypothesis for the above example would be “There is no difference between men and women in terms of map reading ability”

• We then collect some data from a sample of men and women (e.g. 20 of each)

• The question we seek to answer is “What is the probability that the data collected came from a population where men and women have the same map reading abilities?” (i.e. the null hypothesis is correct)
PROBABILITY

• Coolican (2004) gave the following example:
  – “at my local chip shop I am convinced that they save money by giving some people rather thin chips (because they can get more chips from each potato). There are two chip bins under the counter – the owner of the chip shop claims that they contain the same kind of chip, but I suspect they are different. So I tried an experiment. I asked for one bag of chips from each of the chip bins, and measured the width of the chips in each bag”
PROBABILITY

• Coolican (2004) – continued
  – Belief 1 is “The two bins contain chips of an equal average width”
  – Belief 2 is “One bin has thinner chips on average than the other”
  – “In fact I found a very small difference between the average width of the chips in each bag”
PROBABILITY

• We would expect small differences between samples (bags of chips) as things do vary a little – this is simply random variation or “chance”

• What we are looking at is a sufficiently large difference between the samples to be sure the bins (the total population) are actually different

• Otherwise we assume the bins are the same (i.e. the samples are drawn from a single population rather than from two different populations)
PROBABILITY

- The *bins* contain the populations
- The *bags* of the chips are samples
- The **NULL HYOTHESIS (H₀)** - The belief that the two bins contain chips of the same width
  - This is a statement of *no effect* (i.e. the samples are not different)
- The **ALTERNATIVE HYPOTHESIS (H₁)** – The alternative belief is that one bin has thinner chips
  - This is a statement that there is an effect (i.e. the samples are different)
- Ultimately we are interested in making a statement about the population(s) from which the samples are drawn rather than just saying something about samples themselves
NULL HYPOTHESIS

• This is a statement of no difference or no correlation
• It suggests that “nothing is going on”
• For example:
  – You see your best friend’s boyfriend with another girl, doing more than talking. You think to yourself “How likely is it that he would be kissing her if there is nothing going on between them?”
  – **Null Hypothesis** = “There is nothing going on, there is no relationship between them”
  – **Alternative Hypothesis** = “There is something going on between them”
• It is not very likely that he would be kissing her if there was nothing going on, therefore you reject the null hypothesis and accept the alternative hypothesis (i.e. that he is cheating on your friend)
PROBABILITY

• In the previous example, you might have worked out the likelihood that the cheating was real (i.e. you may have felt “fairly certain”)
• In research, we need to be a bit more precise than that
• In order to work out whether a difference is or is not significant we use inferential tests
• These tests allow you to work out how probable it is that a pattern in research data could have arisen by chance or how probable it is that the effect occurred because there is a real difference/correlation in the populations from which the samples were drawn
PROBABILITY - CHANCE

• **Chance** refers to something with no cause, it just happens

• We decide on a probability that we will “risk”

• You cannot be 100% certain that an observed effect was not due to chance but you can state how certain you are

• You may say that you are 99% sure that your friend’s boyfriend is cheating on her (this means that you are fairly confident that you are right, but you may have a little bit of doubt)
PROBABILITY LEVELS

• Psychologists use a level of probability of 95%, which meant that there is a degree of uncertainty.

• This would also mean that there is a 5% chance (probability) of the results occurring if the null hypothesis is true (i.e. there is nothing going on).
  – In other words, a 5% probability that the results would occur even if there was no difference/correlation between the populations from which the samples were drawn.

• The probability of 5% is recorded as $p = 0.05$ (where $p$ means probability). In fact the probability is 5% or less which is written as $p \leq 0.05$.

• In some studies (like replicating a previous study or considering the effects of a new drug on health) psychologists will want to be more certain.

• Researchers use a more stringent probability of 1% or less, which is written as $p \leq 0.01$.

• The chosen value of “$p$” is called the significance level.
TYPE I AND TYPE II ERRORS

• Psychologists generally use the 5% probability level as this is a good compromise between being too strict or too lenient about accepting the null hypothesis (or more formally, a good compromise between a Type I error and a Type II error)

• Type I and II errors concern whether we make a mistake about accepting/rejecting the null hypothesis

• Consider this scenario
  – You are a juror and have to decide whether the person being tried is guilty or innocent
  – See the table on the next slide that shows the four possible outcomes
### TYPE I AND TYPE II ERRORS

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Truth (which we will never know)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Guilty</td>
<td>Not guilty</td>
<td></td>
</tr>
<tr>
<td>Guilty verdict</td>
<td>True positive</td>
<td>False positive (guilt reported, i.e. positive)</td>
<td></td>
</tr>
<tr>
<td>Not guilty verdict</td>
<td>False negative (guilt not detected, i.e. negative)</td>
<td>True negative</td>
<td></td>
</tr>
</tbody>
</table>

- **Nothing is found so it is negative**
- **Something is found so it is positive**
The same can be applied to a research study:

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Truth (which we will never know)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Alternative hypothesis</td>
<td>H₁ true</td>
<td>Null hypothesis</td>
</tr>
<tr>
<td>Reject null hypothesis</td>
<td>True positive</td>
<td></td>
<td>False positive</td>
</tr>
<tr>
<td>Accept null hypothesis</td>
<td>False negative</td>
<td>TYPE II ERROR</td>
<td></td>
</tr>
</tbody>
</table>
TYPE I AND TYPE II ERRORS

• Consider the null hypothesis “There is no difference in the map reading abilities of men and women”

• We conduct a study and find a small difference between men and women

• There are two correct conclusions and two possible errors (as shown in the previous tables)
TYPE I AND TYPE II ERRORS

• In truth there is no difference between men and women but because we use a lenient level of significance (e.g. 10%) we reject the null hypothesis

• This is a **Type I error**, where we reject a null hypothesis that is actually true (e.g. we decide that women are better than men at reading maps whereas, in reality, this is not true)

• **OR** in truth there is a difference between men and women but, because we are using a significance level that is too stringent (e.g. 1%), we accept the null hypothesis when in fact it is false (e.g. women are really better than men at reading maps but we mistakenly accept that there is no difference)

• This is a **Type II error**
SELECTING WHICH STATISTICAL TEST TO USE

• There are a large number of statistical tests that are used by psychologists

• These are divided into **parametric and non-parametric tests**

• Parametric tests are preferred to non-parametric tests as they are more **powerful** (explained later)

• However, they can only be used if certain criteria are met (explained later)

• The tests required in the specification are shown in the decision tree on the next slide along with the criteria to make your choice

NB: matched pairs design counts as a repeated measures (related) design as the two groups of participants are related
SELECTING WHICH STATISTICAL TEST TO USE

SELECTING A STATISTICAL TEST

CAN YOU USE A PARAMETRIC TEST?
- YES
- NO

WAS AN INDEPENDENT GROUPS DESIGN USED?
- YES
- NO

IS THE DATA NOMINAL?
- YES
- NO

IS THE DATA CORRELATIONAL?
- YES
- NO

PEARMON'S r

TEST OF DIFFERENCE
- YES
- NO

UNRELATED t-test
- RELATED t-test

CHI-SQUARED

REPEATED MEASURED SIGN TEST

SPEARMON'S rho

WAS AN INDEPENDENT GROUPS DESIGN USED?
- YES
- NO

MANN-WHITNEY
- REPEATED MEASURED WILCOXON
WHAT DOES “POWERFUL” MEAN

• Parametric statistical tests make calculations using the mean and standard deviation of a data set, whereas non-parametric tests use ranked data, thus losing some of the detail.

• The end result is that parametric tests can detect significance in some situations where non-parametric tests cannot.
### PARAMETRIC CRITERIA

<table>
<thead>
<tr>
<th>CRITERION</th>
<th>EXPLANATION</th>
<th>HOW TO DECIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>The <strong>level of measurement</strong> is interval or better</td>
<td>Levels of measurement were covered in Year 1 (see next slide)</td>
<td>Are the data in categories (nominal) or ordered in some way (ordinal) = non-parametric statistics. Are the intervals between the data truly equal interval = parametric statistics</td>
</tr>
<tr>
<td>The data are drawn from a population that has a normal distribution</td>
<td>Note that it is not the sample that must be normally distributed but the population. A normal distribution is when most items cluster around the mean with an equal number of items above and below the mean</td>
<td>We expect many physical and psychological characteristics to be normally distributed, like height, IQ, shoe size, etc. Therefore you can justify the use of parametric test by saying that the characteristic measured is assumed to be normal. You could also check the distribution of scores to see if they are skewed or not</td>
</tr>
<tr>
<td>The variances of the two samples are not significantly different</td>
<td>The variance is a measure of how spread out a set of data is around the mean. It is the square of the standard deviation</td>
<td>In the case of repeated measures (related samples) any difference in the variances should not distort the result (Coolican, 1996). For independent groups you can check the variances. The variance for one sample should not be more than four times the variance of the other</td>
</tr>
</tbody>
</table>
LEVELS OF MEASUREMENT

• **NOMINAL** – The data are in separate categories, such as grouping your class into people who are tall, medium and short

• **ORDINAL** – Data are ordered in some way, for example, lining up your classmates in order of height

• **INTERVAL** – Data are measured using units of equal intervals, such as when counting correct answers or measuring your classmate’s heights (e.g. 1.41m-1.5m; 1.51m-1.6m; 1.61m-1.7m etc)
USING STATISTICAL TESTS

CALCULATED AND CRITICAL VALUES

• Each statistical test involves taking the data collected in a study and doing some calculations which produce a single number called the test statistic

• In the case of Spearman’s Correlation test, that test statistic is called “rho” whereas the Mann-Whitney test it is “U”

• The rho or U value calculated for any set of data is called the calculated value (as it is based on the calculations made)

• To decide if the calculated value is significant, this figure is compared to another number, found in statistical table

• The value in the statistical table is called a critical value as this is the value that a test statistic must reach in order for the null hypothesis to be rejected
USING STATISTICAL TABLES

• There are different statistical tables for each different statistical test

• To find the appropriate value in the table you need to know:

  1. **Significance Level** selected, usually $p \leq 0.05$ (5%)

  2. **The kind of hypothesis.** It must be determined whether a one-tailed or two-tailed test is required – if the hypothesis was a directional hypothesis, then you would use a one-tailed test; if it was non-directional, you would use a two-tailed test

  3. **Value of N** is the number of participants in the study (N). In studies using independent groups design, there are two values for N (one for each group of participants), which are called $N_A$ and $N_B$. In the case of some tests, such as t-tests and Chi-Squared, you calculate **degrees of Freedom** (df)
TABLES OF CRITICAL VALUES: IMPORTANCE OF R

• Some tests are significant when the calculated value is equal to or **greater than** the critical value; **for others it is the reverse**
• You will need to know which, and you will find it stated underneath each table
• One way to remember is to see if there is a letter R in the name of the test
• If there is an R, then the calculated value should be greater than the critical value (e.g. for Spearman’s, Pearson’s, Chi-Squared and related and unrelated t-test)
• If there is no R (e.g. sign test, Mann-Whitney and Wilcoxon), then the calculated value should be less than the critical value
A LEVEL RESEARCH METHODS

NON-PARAMETRIC TESTS OF DIFFERENCE

MANN-WHITNEY TEST FOR UNRELATED DESIGNS
INVESTIGATION

• One explanation offered to why people fall in love is that it is physiological arousal that gets mislabelled as love

• To test this, White et al (1981) created high and low arousal by asking men to run on the spot for 2 minutes or 15 seconds respectively, and then showed the men a short video of a young woman

• Participants were asked to rate the young woman’s attractiveness
STEP 1: STATE THE HYPOTHESES

• **Alternative hypothesis:** People who run for 2 minutes (high arousal) rate a woman as more attractive than people who run for 15 seconds (low arousal)
  – This is a directional hypothesis and therefore requires a one-tailed test

• **Null hypothesis:** There is no difference in the attractiveness rating scores for the high and low arousal conditions
REASON FOR CHOICE OF TEST

• The hypothesis states a **difference** between two sets of data
• The two sets of data are from separate groups of participants – **unrelated**
• The data are **ordinal** because there are not equal intervals between ratings
STEP 2: PLACE RAW DATA IN A TABLE

- The raw data (collected by a group of students) are the rating scores from each group of participants
- See table on the next slide
<table>
<thead>
<tr>
<th>Column A (High arousal)</th>
<th>Rank</th>
<th>Column B (Low arousal)</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>16</td>
<td>4</td>
<td>4.5</td>
</tr>
<tr>
<td>10</td>
<td>23.5</td>
<td>6</td>
<td>12.5</td>
</tr>
<tr>
<td>8</td>
<td>18.5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>12.5</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>3</td>
<td>2.5</td>
</tr>
<tr>
<td>8</td>
<td>18.5</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>21</td>
<td>6</td>
<td>12.5</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>4</td>
<td>4.5</td>
</tr>
<tr>
<td>10</td>
<td>23.5</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>21</td>
<td>7</td>
<td>16</td>
</tr>
</tbody>
</table>

\[ N_A = 10 \quad R_A = 178.5 \quad N_B = 14 \quad R_B = 121.5 \]
STEP 3: RANK EACH DATA SET

• Rank all data items jointly, following the same steps as in the Wilcoxon test
  – I.e. lowest number receives the rank 1, next lowest = 2, etc
  – If there are two or more of the same number (tied ranks), calculate the final rank by working out the mean of the ranks that would have been given

• As there are so many tied ranks it may be best to work out a frequency on a different page
STEP 4: ADD EACH SET OF RANKS

• Total for Column A is $R_A$ and for Column B is $R_B$
STEP 5: FIND THE CALCULATED VALUE OF U

• U is calculated using whichever total is smaller.
• In this case the smaller value is the total for Column B:
  – \( N_B = 14 \)
  – \( U_2 = R_B - \left[ N_B (N_B + 1) \right] / 2 \)
  – \( U_2 = 121.5 - (14 \times 15) / 2 = 16.5 \)

• \( U = 16.5 \)
STEP 6: IS THE RESULT IN THE RIGHT DIRECTION?

• The predicted direction was that higher arousal would cause higher ratings, so the result is in the right direction
STEP 7: FIND THE CRITICAL VALUE OF U

• The significance level is 5% (p≤0.05)
• The kind of hypothesis is directional and therefore a one-tailed test is used
• $N_A = \text{Number in Column A} = 10$
• $N_B = \text{Number in Column B} = 14$
• Locate the point in the statistical table on p.27 of the textbook where our $N_A$ and $N_B$ value intersect

• The critical value of U = 41
STEP 8: REPORT THE CONCLUSION

• If the calculated value is equal to or less than this critical value, our result is significant.

• In our case it is significant as the calculated value is 16.5 and the critical value is 41.

• As the calculated value is significant (at $p \leq 0.05$), we can reject the null hypothesis and conclude that people with high arousal rate women as more attractive than people with low arousal.
A LEVEL RESEARCH METHODS

NON-PARAMETRIC TESTS OF DIFFERENCE

WILCOXON TEST FOR RELATED DESIGNS
INVESTIGATION

• Research has found that people like things that are familiar

• Zajonc (1968) told participants that he was conducting a study on visual memory and showed them photos of 12 different men (face only), each for two seconds

• At the end, participants were asked to rate how much they liked the 12 different men on a scale from 0-6

• Some photos were shown more often than others, e.g. one photo appeared 25 times, whereas another appeared only once
STEP 1: STATE THE HYPOTHESES

• **Alternative hypothesis:** People rate the more frequently seen face as more likable than the less frequently seen face
  – This is a directional hypothesis and therefore requires a one-tailed test

• **Null hypothesis:** There is no difference in the likability score for faces seen more or less
REASON FOR CHOICE OF TEST

• The hypothesis states a difference between two sets of data
• The two sets of data are pairs of scores from one person – related
• The data are ordinal because there are not equal intervals between ratings
STEP 2: PLACE RAW DATA IN A TABLE

• The raw data (collected by a group of students) are the rating scores for each condition
• There are two items for each participant (see table on the next slide)
<table>
<thead>
<tr>
<th>Participant</th>
<th>Likability for <strong>more</strong> frequently seen faces</th>
<th>Likability for <strong>less</strong> frequently seen faces</th>
<th>Difference</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>9.5</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>omit</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>6.5</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>-1</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>5</td>
<td>-1</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>9.5</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>4</td>
<td>-1</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>9.5</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>6</td>
<td>-2</td>
<td>6.5</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>9.5</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>4</td>
<td>-1</td>
<td>3</td>
</tr>
</tbody>
</table>
STEP 3: FIND THE DIFFERENCE AND RANK

- Calculate the difference between each score
- If the difference is zero, omit this from the ranking and reduce N accordingly
- Rank the differences from low to high, ignoring the signs (- or +)
  - I.e. lowest number receives the rank 1, next lowest = 2, etc
- If there are two or more of the same number (tied ranks), calculate the final rank by working out the mean of the ranks that would have been given
  - In this case 1 occurs 5 times and would have had ranks of 1, 2, 3, 4 and 5, so gets a rank of 3 (1+2+3+4+5 = 15 / 5 = 3)
STEP 4: FIND THE CALCULATED VALUE OF T

• T is the sum of the ranks of the less frequent sign

• In this case the less frequent sign is minus, so we add $3 + 3 + 3 + 6.5 + 3 = 18.5$

• $T = 18.5$
STEP 5: IS THE RESULT IN THE RIGHT DIRECTION?

• The predicted direction was that the more frequent face would be more likable, so the result is in the right direction
STEP 6: FIND THE CRITICAL VALUE OF T

• The significance level is 5% (p ≤ 0.05)
• The kind of hypothesis is directional and therefore a one-tailed test is used
• N value = 11 (as one score was omitted)
• Locate the row in the statistical table on p.26 of the textbook that begins with our N value

• The critical value of T = 13
STEP 7: REPORT THE CONCLUSION

• If the calculated value is equal to or less than this critical value, our result is significant.

• In our case it is not significant as the calculated value is 18.5 and the critical value is 13.

• As the calculated value is not significant (at $p \leq 0.05$), we must accept the null hypothesis and conclude that there is no difference in the likability score of faces seen more or less.
A LEVEL RESEARCH METHODS

PARAMETRIC TESTS OF DIFFERENCE

RELATED T-TEST
INVESTIGATION

• Craik and Tulving (1975) demonstrated that memory for words is better if participants are required to process the words semantically (e.g. FLAME: Does the word mean something hot? Yes/No) rather than just in terms of its physical structure (e.g. FLAME: Is the word in capital letters? Yes/No)

• Semantic processing is considered “deep” processing, whereas processing in terms of physical structure is called “shallow” processing
STEP 1: STATE THE HYPOTHESES

- **Alternative hypothesis:** People remember more words that are semantically processed (deep condition) than words that are processed in terms of physical structure (shallow condition)
  - This is a directional hypothesis and therefore requires a one-tailed test

- **Null hypothesis:** There is no difference in the number of words remembered in the deeply processed or shallowly processed conditions
REASON FOR CHOICE OF TEST

• The hypothesis states a \textit{difference} between two sets of data
• The two sets of data are pairs of scores from one person – \textit{related}
• The data are \textit{interval} because there are equal intervals when counting frequency
• The data fit the criteria for a parametric test: the data are interval, the population is assumed to have a normal distribution and the variances of the samples are the same because they come from the same participants
STEP 2: PLACE RAW DATA IN A TABLE

• The raw data (collected by a group of students) are the number of words recalled for each condition
• There are two items for each participant (see table on the next slide)
<table>
<thead>
<tr>
<th>Column A Deep Processing</th>
<th>Column B Shallow Processing</th>
<th>Difference (d)</th>
<th>d²</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>18</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>18</td>
<td>19</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>22</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>16</td>
<td>-4</td>
<td>16</td>
</tr>
<tr>
<td>16</td>
<td>13</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>16</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>24</td>
<td>20</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>21</td>
<td>14</td>
<td>7</td>
<td>49</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

Mean = 17.9  
Mean = 16.4  
\[ \sum d = 15 \]  
\[ \sum d^2 = 171 \]
**STEP 3: FIND THE CALCULATED VALUE OF T**

- The number of participants (N) was 10

\[ t = \sqrt{\frac{\Sigma d^2 - (\Sigma d)^2}{N \cdot (N-1)}} \]

- \[ t = \sqrt{\frac{171 - 225}{10}} \]

- \[ t = \sqrt{\frac{148.5}{90}} \]

- The Calculated Value of \( t = 1.285 \)

**NOTE:** You may be asked to calculate this in the exam using a calculator, but the formula will be given in the exam.
STEP 4: IS THE RESULT IN THE RIGHT DIRECTION?

• The predicted direction was that deep processing would produce a higher score, so the result is in the right direction
STEP 5: FIND THE CRITICAL VALUE OF T

• The significance level is 5% (p ≤ 0.05)
• The kind of hypothesis is directional and therefore a one-tailed test is used
• df value = N – 1 = 9
• Locate the row in the statistical table on p.28 of the textbook that begins with our df value

• The critical value of t = 1.833
STEP 6: REPORT THE CONCLUSION

• If the calculated value is equal to or **greater** than this critical value, our result is significant
• In our case it is not significant as the calculated value is 1.285 and the critical value is 1.833
• As the calculated value is not significant (at p≤0.05), we must accept the null hypothesis and conclude that there is no difference in the number of words remembered in the deeply processed or shallowly processed conditions
A LEVEL RESEARCH METHODS

PARAMETRIC TESTS OF DIFFERENCE

UNRELATED T-TEST
INVESTIGATION

• Piliavin et al (1969) investigated helping behaviour, and looked specifically at whether people were quicker at offering help in an emergency situation to a man with a cane or a man who appeared to be drunk

• The emergency situation was when a person appeared to collapse on an underground train

• Participants were randomly assigned to condition A (man with a cane) or condition B (man who appears to be drunk)
STEP 1: STATE THE HYPOTHESES

• **Alternative hypothesis:** The speed of offering help in an emergency situation is faster when the victim is carrying a cane than when appearing drunk
  – This is a directional hypothesis and therefore requires a one-tailed test

• **Null hypothesis:** There is no difference in the speed of offering help to victims with a cane or appearing drunk
REASON FOR CHOICE OF TEST

• The hypothesis states a *difference* between two sets of data
• The two sets of data are pairs of scores from separate groups of participants – *unrelated*
• The data are *interval* because there are equal intervals when counting frequency
• The data fit the criteria for a parametric test: the data are interval, the populations are assumed to have a normal distribution and the variances of the samples are assumed to be the same, as the participants were randomly assigned to conditions
STEP 2: PLACE RAW DATA IN A TABLE

• The raw data are the number of seconds it took for help to be offered
• See table on the next slide
<table>
<thead>
<tr>
<th>Column A Time for a man with cane to be helped</th>
<th>Column B Time for a man appearing drunk to be helped</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>74</td>
</tr>
<tr>
<td>34</td>
<td>96</td>
</tr>
<tr>
<td>49</td>
<td>83</td>
</tr>
<tr>
<td>74</td>
<td>80</td>
</tr>
<tr>
<td>61</td>
<td>40</td>
</tr>
<tr>
<td>30</td>
<td>22</td>
</tr>
<tr>
<td>52</td>
<td>97</td>
</tr>
<tr>
<td>83</td>
<td>121</td>
</tr>
<tr>
<td>110</td>
<td>92</td>
</tr>
<tr>
<td>43</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>85</td>
</tr>
</tbody>
</table>

$N_A = 10$  $N_B = 12$

Mean = 59.2  Mean = 86.75
STEP 3: FIND THE CALCULATED VALUE OF T

- The formula for the unrelated t test is very complex and would not be used in an exam.
- There are many online calculators that will calculate this for you (e.g. [www.socscistatistics.com/tests/studenttest/](http://www.socscistatistics.com/tests/studenttest/)).
- The result of this calculation is 1.882.
- So the Calculated Value of t = 1.882.
STEP 4: IS THE RESULT IN THE RIGHT DIRECTION?

• Yes, the results are in the predicted direction
STEP 5: FIND THE CRITICAL VALUE OF T

• The significance level is 5% (p≤0.05)
• The kind of hypothesis is directional and therefore a one-tailed test is used
• df value = (N_A + N_B) − 2 = 20
• Locate the row in the statistical table on p.28 of the textbook that begins with our df value (this is the same table used for the Related t Test)

• The critical value of t = 1.725
STEP 6: REPORT THE CONCLUSION

• If the calculated value is equal to or greater than this critical value, our result is significant.
• In our case it is significant as the calculated value is 1.882 and the critical value is 1.725.
• As the calculated value is significant (at $p \leq 0.05$), we are able to reject the null hypothesis.
• But, if the result was in the wrong direction, we can’t accept the alternative hypothesis and so have to accept the null hypothesis.
• We therefore can’t draw any conclusion (and we cannot change the hypothesis).
A LEVEL RESEARCH METHODS

TESTS OF CORRELATION

A PARAMETRIC TEST:
PEARSON’S R
INVESTIGATION

• We have previously looked at inter-observer reliability (p.16 of the textbook), and how correlational tests are used to determine the degree to which observations from two observers are related
STEP 1: STATE THE HYPOTHESES

• **Alternative hypothesis:** The observations from observer A are positively correlated to the observations from observer B
  – This is a directional hypothesis and therefore requires a one-tailed test

• **Null hypothesis:** There is no correlation between observations made by observer A and observer B
REASON FOR CHOICE OF TEST

• The hypothesis states a correlation between two sets of data
• The two sets of data are pairs of scores that are related
• The data are interval because they are counting the number of observations
• The data fit the criteria for a parametric test: the data are interval, the populations are assumed to have a normal distribution and the variances of the samples are assumed to be the same, as the participants are related.
STEP 2: PLACE RAW DATA IN A TABLE

• The observations from two observers (A and B) are placed in a raw data table
• See table on the next slide
<table>
<thead>
<tr>
<th>Behaviour</th>
<th>Observer A</th>
<th>Observer B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hits</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Touches</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Cuddles</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Sits next to</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Talks</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Running about</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Crying</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Laughing</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Smiling</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
STEP 3: FIND THE CALCULATED VALUE OF R

- The formula for Pearson’s r is very complex and would not be used in an exam
- There are many online calculators that will calculate this for you (e.g. www.socscistatistics.com/tests/pearsons)

- The result of this calculation is 0.470
- So the Calculated Value of r = 0.470
STEP 4: IS THE RESULT IN THE RIGHT DIRECTION?

• The predicted direction was a positive correlation, so the result is in the right direction
STEP 5: FIND THE CRITICAL VALUE OF T

• The significance level is 5% (p≤0.05)
• The kind of hypothesis is directional and therefore a one-tailed test is used
• df value = N – 2 = 7 (number of behaviour categories)
• Locate the row in the statistical table on p.31 of the textbook that begins with our df value

• The critical value of r = 0.582
STEP 6: REPORT THE CONCLUSION

• If the calculated value is equal to or **greater** than this critical value, our result is significant

• In our case it is not significant as the calculated value is 0.470 and the critical value is 0.582

• As the calculated value is not significant (at \( p \leq 0.05 \)), we must accept the null hypothesis and conclude that there is no correlation between observations made by observer A and observer B
A LEVEL RESEARCH METHODS

TESTS OF CORRELATION

A NON-PARAMETRIC TEST:
SPEARMAN’S RHO
INVESTIGATION

• Brosnan (2008) found that boys with smaller finger length ratio between their index and ring fingers were more likely to have a talent in Maths

• The explanation is related to the effects of testosterone which is associated with reduced finger length ratio and increased numeracy skills
STEP 1: STATE THE HYPOTHESES

- **Alternative hypothesis**: The finger length ratio between index finger and ring finger is negatively correlated to numeracy skills in boys
  - This is a directional hypothesis and therefore requires a one-tailed test

- **Null hypothesis**: There is no correlation between finger length ratio and numeracy in boys
REASON FOR CHOICE OF TEST

• The hypothesis states a *correlation* between two sets of data
• The two sets of data are pairs of scores from one person – *related*
• The data are *ordinal* because numeracy skills are measured using a test and may not have equal intervals between scores
STEP 2: PLACE RAW DATA IN A TABLE

• The raw data are the finger length ratio and numeracy score for each person
• See calculation table on the next slide
<table>
<thead>
<tr>
<th>Column A</th>
<th>Rank A</th>
<th>Column B</th>
<th>Rank B</th>
<th>Difference between Rank A and Rank B (d)</th>
<th>d²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finger length ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.026</td>
<td>10</td>
<td>8</td>
<td>2.5</td>
<td>7.5</td>
<td>56.25</td>
</tr>
<tr>
<td>1.000</td>
<td>5.5</td>
<td>16</td>
<td>9</td>
<td>-3.5</td>
<td>12.25</td>
</tr>
<tr>
<td>1.021</td>
<td>9</td>
<td>10</td>
<td>5</td>
<td>4.0</td>
<td>16</td>
</tr>
<tr>
<td>0.991</td>
<td>4</td>
<td>9</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.984</td>
<td>3</td>
<td>15</td>
<td>8</td>
<td>-5.0</td>
<td>25.0</td>
</tr>
<tr>
<td>0.975</td>
<td>1</td>
<td>14</td>
<td>7</td>
<td>-6.0</td>
<td>36.0</td>
</tr>
<tr>
<td>1.013</td>
<td>7</td>
<td>12</td>
<td>6</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>1.018</td>
<td>8</td>
<td>8</td>
<td>2.5</td>
<td>5.5</td>
<td>30.25</td>
</tr>
<tr>
<td>0.982</td>
<td>2</td>
<td>17</td>
<td>10</td>
<td>-8.0</td>
<td>64.0</td>
</tr>
<tr>
<td>1.000</td>
<td>5.5</td>
<td>5</td>
<td>1</td>
<td>4.5</td>
<td>20.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Σ d² = 261.0</strong></td>
<td></td>
</tr>
</tbody>
</table>
STEP 3: FIND THE CALCULATED VALUE OF RHO

• The formula is shown below

• In order to use this formula you need to:
  – Rank data in Column A from low to high and do the same for data in Column B
  – Find the difference between each pair of ranks
  – Square this difference and add the column up (Σ d²)

• \[ \rho = 1 - \frac{6 \Sigma d^2}{N (N^2 - 1)} \]

• \[ \rho = 1 - \frac{6 \times 261.0}{10 \times (100 - 1)} \]

• \[ \rho = 1 - \frac{1566}{990} \]

• So the Calculated Value of \( \rho \) = - 0.58
STEP 4: IS THE RESULT IN THE RIGHT DIRECTION?

• The predicted direction a was negative correlation, so the result is in the right direction
STEP 5: FIND THE CRITICAL VALUE OF T

• The significance level is 5% (p≤0.05)
• The kind of hypothesis is directional and therefore a one-tailed test is used
• N value = 10
• Locate the row in the statistical table on p.30 of the textbook that begins with our N

• The critical value of rho = 0.564
STEP 6: REPORT THE CONCLUSION

• If the calculated value is equal to or greater than this critical value, our result is significant
• In our case it is significant as the calculated value is -0.58 and the critical value is 0.564
  – Even though the calculated value is negative, it is only the value (not the sign) that is important when comparing to the critical value (although the sign does tell you if the correlation is positive or negative)
• As the calculated value is significant (at p≤0.05), we can reject the null hypothesis and conclude that finger length ratio between index finger and ring finger is negatively correlated to numeracy skills in boys
A LEVEL RESEARCH METHODS

CHI-SQUARED TEST ($X^2$)
INVESTIGATION

• Gilligan and Attanucci (1988) found that women were more likely to make moral decisions based on an ethic of care than an ethic of justice, whereas men made moral decisions based the other way round
STEP 1: STATE THE HYPOTHESES

- **Alternative hypothesis:** There is an association between gender and the kind of moral decisions made (based on care or justice)

OR

- **Alternative hypothesis:** There is a difference between men and women in terms of the basis for making moral decisions
  - Both of the above are non-directional hypotheses and therefore require a two-tailed test

- **Null hypothesis:** There is no association/difference between men and women in terms of the basis for making moral decisions
REASON FOR CHOICE OF TEST

• The hypothesis states a *difference/association* between two sets of data
• The data in each cell *independent*
• The data are *nominal* because each person belongs to one of the four categories
STEP 2: PLACE RAW DATA IN A TABLE

• The raw data (collected by a group of students) have been placed in a 2 X 2 contingency table (see next slide)
• The number of people in each cell is recorded and then totals for each row and column are calculated
<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethic of Justice</td>
<td>5</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>(cell A)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethic of Care</td>
<td>10</td>
<td>9</td>
<td>19</td>
</tr>
<tr>
<td>(cell C)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>15</td>
<td>21</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
STEP 3: FIND THE OBSERVED/CALCULATED VALUE OF $X^2$

• The expected frequencies are calculated following the steps in the table on p.32 of the textbook
• This may look complicated but it isn’t
• Add all the values in the final column go get the observed/calculated value of $X^2$

• From the table, the observed/calculated value of $X^2$ is 1.984
STEP 4: FIND THE CRITICAL VALUE OF $X^2$

- The significance level is 5% ($p \leq 0.05$)
- The kind of hypothesis is non-directional and therefore a two-tailed test is used
- $df$ value = ($\text{number of rows} - 1$) $\times$ ($\text{number of columns} - 1$)
- $df$ value = $(2-1) \times (2-1) = 1$
- Locate the row in the statistical table on p.32 of the textbook that begins with our $df$ value
- The critical value of $X^2 = 3.84$
STEP 5: REPORT THE CONCLUSION

• If the calculated value is equal to or greater than this critical value, our result is significant.
• In our case it is not significant as the calculated value is 1.984 and the critical value is 3.84.
• As the calculated value is not significant (at p≤0.05), we must accept the null hypothesis and conclude that there is no association between men and women in terms of the basis for making moral decisions.
A LEVEL RESEARCH METHODS

REPORTING INVESTIGATIONS

JOURNAL ARTICLES
• Studies are written up and published in peer-reviewed academic journals for everyone to read.
• You can access many original reports online and you will see that they are almost always organised into the following sections:
  – Abstract
  – Introduction
  – Method
  – Results
  – Discussion
  – References
ABSTRACT

• This is a summary of the study covering the aims, hypothesis, the method (procedures), results and conclusions (including implications of the current study)

• This is usually about 150-200 words in length and allows the reader to get a quick picture of the study and its results
INTRODUCTION

• This begins with a review of previous research (theories and studies) in the same area, so the reader knows what other research has been done and understands the reasons for the current study.

• The focus of this research review should lead logically to the study to be conducted so the reader is convinced of the reasons for this particular research.

• The introduction should be like a funnel – starting broadly and narrowing down to the particular research hypothesis.

• The researcher states their aims, research prediction and/or hypothesis.
METHOD

• This section contains a detailed description of what the researcher did, providing enough information for replication of the study
  – **Design**, e.g. “repeated measures” or “covert observation”. Design decisions should be justified
  – **Participants** – Information about sampling methods and how many participants took part and their details (e.g. age, job, etc)
  – **Apparatus/Materials** – Description of any materials used
  – **Procedures**, including standardised instructions, the testing environment, the order of events etc
  – **Ethics** – Significant ethical issues may be mentioned, as well as how they were dealt with
RESULTS

• Details are given about what the researcher found, including:
  – **Descriptive Statistics** – Tables and graphs showing frequencies and measures of central tendency and dispersion
  – **Inferential Statistics** (statistical tests) are reported, including calculated values and significance level
    • Selecting the right test and justifying it
  – In the case of qualitative research, categories and themes are described along with examples within these categories
DISCUSSION

• In this section the researcher aims to interpret the results of the study and consider their implications for future research as well as suggesting real-world applications

  – Summary of the results – The results are reported in brief and some explanation given about what these results show

  – Relationship to previous research – The results of the study are discussed in relation to the research reported in the introduction and possibly other research not previously mentioned

  – Consideration of Methodology – Criticisms may be made of the methods used in the study, and improvements suggested

  – Implications for psychological theory and possible real-world applications

  – Suggestions for future research
REFERENCES

• The full details of any journal articles or books that are mentioned in the research report are given, just like the ones at the back of the textbooks.

• The format for journal articles generally is:
  – Author’s name(s), date, title of article, journal title, volume (issue number), page numbers.

• If it is a book it is:
  – Author’s name(s), date, title of book, place of publication, publisher.