INTERNAL ASSESSMENT

Internal assessment is an integral part of the IB psychology course. It is responsible for 25% of your marks at SL and 20% at HL. This component is internally assessed by your teacher and externally moderated by the IB at the end of the course.

The aim of the internal assessment component is for you to demonstrate the application of the skills and knowledge that you have acquired by studying psychology, with sufficient time to plan and implement your project without the stress and time constraints associated with a regular examination situation.

You need to select a published piece of psychological research and replicate it with the aim of investigating the underlying psychological model or theory. There are certain limitations, both methodological and ethical, to what studies you can replicate. The result of your work will be a written report documenting all stages of the investigation. You will be required to conduct the experiment, collect data, analyse data in line with your aim and hypothesis, and report and discuss the findings.

The internal assessment component relies heavily on your knowledge of the experiment as a method of research in psychology. The skills that you used when evaluating experimental studies discussed throughout the book will come in handy as you carry out your own investigation and analyse the findings. The discussion of the experiment in Unit 1 on research methodology is also useful.

To analyse the findings in your experiment you will need to apply some simple statistical tests. You are free to use software to do that, but you need to understand which statistical tests to choose, what output values to look at and how to interpret the results. To make this process more transparent, we are walking you through manual calculations of the most popular statistical tests. If you understand how to calculate manually, you will have no difficulty with using software properly.

What you will learn in this unit

- Overview of the requirements for internal assessment
- Planning the investigation
  - Choosing the study
  - Simplification and modification
  - Examples: Bransford and Johnson (1972); Loftus and Palmer (1974)
- Writing the introduction
- Writing the exploration
- Conducting the analysis
  - Descriptive statistics: levels of measurement; measures of central tendency; measures of dispersion; graphing the results
  - Inferential statistics: overview; unrelated t-test; related t-test; Mann–Whitney U test; Wilcoxon signed-rank test
- Writing the evaluation
- References and appendices

sample of 1,500 rural children aged 9–13 years and annual psychiatric assessments. This continued for eight years (1993–2000). One quarter of the sample were Native American, the rest were predominantly white. The naturally occurring variable was a casino that opened on the Indian reservation. This happened halfway through the study and it gave every Native American an income supplement that increased annual incomes of non-Indian families were unaffected. The increase of income changed the poverty status of some of the Native American families: 14% of the study's families moved out of poverty (the "ex-poor" group), 53% remained poor ("persistently poor"), while the remaining 32% were the "never poor" group. Psychiatric symptoms (assessed against DSM-IV) were compared in the never poor, persistently poor and ex-poor children in the four years before and after the casino opened.

Results showed that for the four years before the casino opened the persistently poor and ex-poor children had more psychiatric symptoms than never poor children. However, in the four years after the casino opened the psychiatric symptoms among the ex-poor children dropped to the level of never poor children. At the same time levels of psychiatric symptoms among persistently poor children remained high.

This decrease of symptoms for ex-poor children was not universal: externalizing behaviours (such as conduct disorder and aggression) were affected, but internalizing behaviours such as anxiety and depression were not.

Similar patterns were observed in non-Native American families that moved out of poverty due to reasons other than the casino opening.

How does this natural experiment further our understanding of poverty as a factor that may or may not influence children's development by itself, putting aside the influence of other associated variables? There are two competing hypotheses as to what may cause symptoms of psychiatric disorders in children: poverty or family characteristics. Family characteristics (such as psychiatric symptoms found in parents and parents' education) remained constant throughout the course of the study. **Financial status** of the family was the variable that changed. Therefore, we may conclude that poverty has an effect on the child's development over and above relatively stable family characteristics, at least for some disorders (such as conduct disorder and aggression).

Discussion

To what extent do these two studies—Dickerson and Popli (2016) and Costello *et al* (2003)—solve the problem of separating effects of poverty on cognitive and social development from all other associated effects?

Of course it is not possible to isolate all accompanying effects completely. For example, changing financial circumstances may lead to increased time interacting with children in the family or to better access to quality educational materials. However, the research findings allow us to come closer to pinpointing the exact developmental outcomes that depend on the exact isolated factors. This knowledge can potentially enable us to implement better intervention programmes that would mitigate the negative effects of poverty on children's development.

Psychology in real life

Now we know that poverty, together with poverty-related factors, can influence a child's cognitive and social development. Moreover, we know that the timing of poverty episodes plays a huge role (the earlier these episodes occur, the more risky they are). The question is, what can we do about it? Given that we cannot just help all affected families financially, can we use whatever limited resources are available to reduce the risks associated with poverty experienced in early childhood? This is the task for non-governmental organizations so they need to be smart, they need to target resources to areas that are most crucial for child development, and the interventions need to be evidence-based.

An example of a programme that aims to reduce the effects of poverty on the development of children on a national level is Head Start. Currently Head Start is a programme of the US Department of Health and Human Services. It was launched in 1965, originally as a summer school programme that taught low-income children in a few weeks the basics so that they could successfully start elementary school. Now it provides comprehensive early education, health, nutrition and parent involvement services.

Review the website of Head Start programme. Explore the range of services it offers.

http://www.nhsa.org/

Can you suggest any additional services that you think might have a positive impact?

9 Overview of the requirements for internal assessment

Internal assessment in psychology is compulsory for students at both SL and HL with the same requirements for both levels. You have to carry out a replication of an experiment and report the results.

Choice of topic

The topic of your investigation can be taken from any area of psychology as long as it meets the methodological, ethical and other requirements outlined below.

- The theory, model or the research study upon which the investigation is based must appear in a peer-reviewed journal.
- This must be an experiment. As you know, that means that at least one independent variable (IV) must be manipulated, at least one dependent variable (DV) must be accurately measured and the potential confounding variables must be controlled.
- Quasi-experiments are not suitable for internal assessment. In quasi-experiments the IV is not manipulated by the researcher but occurs naturally. Examples include:
  - gender
  - age
  - native language
  - culture
  - socio-economic status
  - left- or right-handedness.
- You cannot conduct experiments involving the following elements:
  - placebos
  - ingestion or inhalation (for example, food, drink, smoking)
  - deprivation (for example, sleep, food).
- Only one IV must be used in the experiment.
- In the study that your investigation is based on, the IV will have two or more levels (conditions). You may use all the levels used in the original study or limit the number of conditions. However, it is strongly recommended to simplify the original study and only use two levels of the IV, otherwise it becomes increasingly difficult to present a strong analysis within the permitted word count.
- Only one DV must be used. Operationalization of this variable may be based on the original study or adapted to better suit your context (for example, you may alter the type of measurement taken). You can even use a different DV altogether, but the link between the original study and your experiment must then be clearly justified.

Word count

The word count for the report is between 1,800 and 2,200 words. The report includes the following components:

- Introduction
- Exploration
- Analysis
- Evaluation
- References
- Appendices

Ethical guidelines

Ethical guidelines must be adhered to throughout the process. These include, but are not limited to, the following.

- It is not permitted to carry out a study that creates anxiety, pain, stress or discomfort in the participants. This is why you should rule out experiments involving, for example, unjustified deception, conformity or obedience.

- Partial deception may be allowed on these conditions.
  - Participants' knowledge of the real aim of the experiment would fundamentally affect the outcome.
  - Deception results in no harm to the participants.
  - At the end of the experiment participants are fully debriefed.
  - At the end participants are given the right to withdraw their data.

↓ Must be stated in the standardized debrief

- Studies involving involuntary participation or invasion of privacy are not allowed. All participants must be informed before the experiment that they have the right to withdraw at any time. — In informed consent
- Informed consent must be gained from all participants in writing.
- Children younger than 12 years old must not be used as participants. For participants aged 12 to 16 years informed consent must be obtained from parents or guardians. — I will provide a form.
- Non-human animals must not be used.

State that you got this consent in your paper

Planning the investigation

Choosing the study

The starting point of your investigation is to select a research study that you want to replicate. Apart from the study meeting all the requirements outlined above, these points are essential.

- You have found the original article in a peer-reviewed journal (or a secondary source which gives sufficient information about the study, including its procedural details and numerical findings).
- You have identified the theory or the model upon which the study is based. Very often this theory or model will be discussed in the article itself as the authors provide theoretical justification for the research.

Simplification and modification

The original study will usually have to be simplified for the purposes of your internal assessment. You may also make a small modification to the procedure.

The simplification of the original study will usually involve one or several of:

- reducing the number of IVs (you must have only one)
- reducing the number of DVs (only one is necessary)
- reducing the number of levels of the IV, that is, the groups or conditions
- simplifying the sampling procedure—you will not have access to the same resources that the original researchers did. (IB students often use other students from the same school as their participants.) – Like us

The modification of the original study must be a minor change so that it is still a replication, not a different study, and a direct comparison is possible with the original findings. This means that your replication must test the same hypothesis and link to the background theory, model or study in the same way. Consequently, the nature of the IV cannot be changed either. Examples of modifications that would be acceptable are:

- the way you measure the DV
- a completely changed DV, as long as the link to the background theory or model is retained
- small details of the procedure such as an additional control variable
- the nature of the sample.

Examples

Let's use two examples of studies discussed in this book: Bransford and Johnson (1972) and Loftus and Palmer (1974) (see Unit 3 on the cognitive approach to behaviour). Both of them are popular choices among IB students, so you are encouraged to look for other studies that would be more in line with your interests.

Psychology in real life

Below you will find the details of the original studies published in peer-reviewed journals. Both papers are available online for free and you can find them using Google Scholar or similar search engines. Look through them one more time before reading on.

- Example A: Bransford, JD and Johnson, MK. 1972. "Contextual prerequisites for understanding: Some investigations of comprehension and recall". *Journal of Verbal Learning and Verbal Behavior*. Vol 11. Pp 716–726. https://tinyurl.com/kp479dh

- Example B: Loftus, EF and Palmer, JC. 1974. "Reconstruction of automobile destruction: An example of the interaction between language and memory". *Journal of Verbal Learning and Verbal Behavior*. Vol 13. Pp 585–589. https://tinyurl.com/y772kup4

Example A

Reading the original Bransford and Johnson paper (1972), we find the following details.

Theory or model upon which the study is based

In the introduction to their paper, Bransford and Johnson discuss how prior knowledge is necessary for the meaningful processing of information. This links to the role of schema in memory at the stage of information encoding. The theory upon which the study is based is schema theory, but more specifically a model that postulates the influence of prior knowledge on comprehension and memory of linguistic material.

Variables

There are four separate experiments reported in this paper. We will be using the first one because it is the one discussed in detail in this book (see Unit 3).

The IV in this experiment is the availability of prior knowledge. This is operationalized as a picture giving the context of the text passage.

The IV has five levels:

1. no context
2. no context, passage read twice
3. full context before
4. partial context before
5. full context after.

There are two DVs:

1. comprehension (subjects were asked to rate how difficult the passage was for comprehension, on a scale from 1 to 7)
2. recall (subjects were given seven minutes to recall all the details they could; researchers later counted the number of correctly recalled details).

Sample and design

This experiment uses an independent measures design, with different groups of participants randomly assigned to different levels of the IV. The sample consists of 50 male and female high school students who volunteered to participate in the experiment (no further information about the sample is given in the paper).

What can be simplified and modified for IA purposes

Consider the following.

- First, you must choose one of the DVs. You are allowed to use a different DV altogether, but in this context it is hard to think of something that links to the background theory or model in the same way. Suppose you have chosen recall.
- You are also allowed to change the way the DV is measured. In the original experiment, participants were given a blank sheet of paper and asked to recall the passage as closely as they could in seven minutes. Fourteen idea units were designated beforehand, and the recall protocols were scored by two independent judges against this list of idea units. Any differences were resolved by the third judge. Paraphrases were allowed. However, the list of idea units themselves is not given in the paper. So you would probably be justified in making the following modifications: use one judge only for the sake of simplicity, and design your own list of idea units based on the original text passage.
- Five levels of the IV is also too many, so you might want to reduce the number of levels to two, for example, "no context" and "full context before".

Example B

Reading the original Loftus and Palmer paper (1974) we find the following details.

Theory or model upon which the study is based

This study is based on the theory of reconstructive memory: post-event information may change one's memory of the event. The significance of this theory is in its link to real-life situations of eyewitness testimony (see Unit 3).

Variables

Two experiments are reported in the paper. We will focus on the first one. There is one IV and one DV. Post-event information (the IV) is operationalized as the emotional intensity of the key verb in the question ("smashed", "collided", "bumped", "hit", "contacted"). Memory of the event (the DV) is operationalized as the estimate of the speed with which the cars were moving, the answer to the question, "About how fast were the cars moving when they hit (smashed into, collided with, etc.) each other?"

Sample and design

There are five groups of participants in the experiments (independent measures design). The total sample consisted of 45 students (no further information about the sample is given in the paper).

What can be simplified or modified for IA purposes

Consider the following.

- You should reduce the number of levels of the IV (and the number of groups). For example, you could use only the conditions "smashed" and "hit".
- Since it is a replication, for the purposes of comparing your results with the original study it makes sense to use the same DV measured in the same way (you may have a different opinion and that is fine as long as you justify it).
- You may need to modify the procedure, however, because the original videos used in the experiment are not available.

Writing the introduction

The *Psychology guide* provides the assessment criteria against which your work will be assessed. For the introduction (assessed against criterion A), the top markband for 5–6 marks states:

> "The aim of the investigation is stated and its relevance is explained.
>
> The theory or model upon which the student's investigation is based is described and the link to the student's investigation is explained.
>
> The independent and dependent variables are stated and operationalized in the null or research hypotheses."

Aim and relevance of the investigation

As we can see above, the first part of the markband descriptor relates to the aim of the investigation and its relevance.

The way the aim is formulated sets the focus for the whole investigation, so choose your wording carefully. In many published research papers the aim is stated explicitly, and since this is a replication you would only need to adapt it to the context of your investigation.

With rare exceptions the aim is a sentence that brings together three elements:

- the IV
- the DV
- a statement of cause-effect relationship.

Note that the IV and the DV in the aim should be presented as constructs, not operationalizations. For example, the IV in Loftus and Palmer (1974) was "post-event information", and it was operationalized through the emotional intensity of the verb in the leading question. Similarly, the DV was "memory", operationalized as a speed estimate. The aim will usually connect the two constructs in a statement of causal relationship, for example, "To investigate the effect of post-event information on memory of the event in an eyewitness situation".

Now that you have stated the aim, you should explain its relevance. This is a brief explanation as to why this research is worth doing.

Theory or model that the investigation is based upon

The second part of the markband descriptor focuses on a description of the theory or model upon which the investigation is based and an explanation of the link to the investigation.

This requires a clear description of the background theory or model and the replicated experiment. Reading the original paper will help you greatly in understanding the connections to the background theory because authors will typically make links to theory when they formulate their hypothesis as well as later when they discuss their findings. In Loftus and Palmer (1974) you will probably focus on the theory of reconstructive memory, and in Bransford and Johnson (1972) on schema theory and the idea that prior knowledge affects comprehension and memory of new information.

You will also need to describe relevant details of the original experiment. Relevance is determined by the aim of your investigation. To make a comparison of your findings with the original findings more effective, you will need to include information about the sample, experimental design, essential details of the procedure, controls, and also ideally state the numerical results from the original study, both descriptive and inferential statistics. Use whatever information is available in the paper, but only mention what is absolutely relevant to your hypothesis.

After providing details of the original study you will also explain the nature of your simplification and modification (if any). If modification has been done, this needs to be justified briefly.

Stating the variables and formulating the hypothesis

The third part of the markband descriptor requires that both the IV and DV are stated (as constructs)

and operationalized. The research hypothesis is a prediction that you are making about the cause–effect relationship between the two variables.

It is recommended to have a set of three statements.

Theoretical prediction

This is the prediction using theoretical terms (constructs) rather than operationalizations. It is linked directly to the background theory or model, for example, "Post-event information will influence memory of an event", or "Presence of contextual information will enhance comprehension and memory of a text". The theoretical prediction is similar to the aim.

Operationalized research hypothesis

This is the theoretical prediction "translated" in operationalized terms. The research hypothesis should be formulated in a way that makes the following points clear:

- what research design is used—comparing different groups of participants or comparing different conditions in the same group
- how the variables are measured
- what is the direction of relationship.

Regarding the last point, hypotheses can be **directional** or **non-directional**. The former predicts the direction, for example, "X will be higher than Y". The latter postulates a relationship, but does not specify its direction, for example, "X will be different from Y".

Using the Loftus and Palmer (1974) example, the operationalized research hypothesis may be something like, "Participants who are asked a question with a verb of higher emotional intensity ('smashed') will report higher speed estimates than participants who are asked a question with a verb of lower emotional intensity ('contacted')". Note that this statement uses operationalized variables; it is easy to tell that different groups of participants will be allocated to different conditions, and the hypothesis is directional.

Null hypothesis

A **null hypothesis** is a statement that is opposite to the operationalized research hypothesis. Usually it expresses the idea that "X will not be different from Y". For "X is independent of Y", for example, "Speed estimates reported by participants who were asked a question with a verb of higher emotional intensity ('smashed') will not be different from speed estimates reported by participants who were asked a question with a verb of lower emotional intensity ('contacted')".

Why do we even need a null hypothesis? In accordance with the principles of falsification of Karl Popper (remember your TOK classes), the proper scientific way to test theories is to try to refute them. If you try to refute a theory but fail, it increases your trust in a theory. Accordingly, we test the null hypothesis. If it gets rejected, we accept the research hypothesis.

Writing the exploration

Looking again at the *Psychology guide* and the assessment criteria against which your work will be assessed, for the exploration (assessed against criterion B), the top markband for 3–4 marks states:

- The research design is explained.
- The sampling technique is explained.
- The choice of participants is explained.
- Controlled variables are explained.
- The choice of materials is explained.

Explaining the research design

As you saw in Unit 1 on research methodology, experiments have three types of design: independent measures, dependent measures and matched pairs. You need to correctly identify the type of design you are using and explain it. Both Loftus and Palmer (1974) and Bransford and Johnson (1972) used independent measures design where participants were randomly allocated into groups.

Note the following points.

- When you use the independent measures design, allocation into groups must be random. You need to think of a procedure that ensures sufficient randomness of allocation, otherwise there would be incorrect implementation of the experimental design. For example, let's say that for a particular male participant you toss a coin and this determines that he has to be allocated to the experimental group. Then you decide to allocate him to the control group anyway to make the male/female ratio in the two groups more equal. This defeats the purpose of the independent measures design.
- When you use the repeated measures design, order effects are inevitable, so you must use counterbalancing. This means that you will end up having two groups of participants anyway, differing in the order of the experimental conditions they are exposed to (AB and BA). Note that when it comes to result analysis, you will still be comparing conditions, not groups. You will compare A from the first group and A from the second group (collated together) to B from the first group and B from the second group (also collated together).

So, what does the markband descriptor mean when it says you need to explain the research design?

First, explain what features of your experimental design allow you to identify it as a particular type. For example, you might say that two separate groups of participants were used with random allocation into the groups, therefore the design was independent measures.

Second, explain why you have chosen this design over the alternatives. Depending on the aim and the context of your investigation, there can be various reasons that could potentially drive this choice. Here are just a few of them.

- Repeated measures design may be preferred because it minimizes the effect of participant variability (initially existing differences between participants). However, order effects are an issue.
- Independent measures design may be preferred because it is not susceptible to order effects: participants only take part in one experimental condition, so they cannot practise, do not get too tired, and it is harder for them to work out the real aim of the study if you have used deception. However, participant variability could be an issue.
- You may choose the experimental design that was used in the original study because it is important for you to compare your findings to the original findings as precisely as possible.

Explaining the sampling technique and the choice of participants

As you will remember, several sampling techniques may be used in experiments: random, stratified, self-selected and opportunity sampling. Start with identifying your target population—this is the group of people you will generalize your findings to. If there are reasons to believe that the phenomenon you are studying is universal

and does not differ with age, education, culture or other demographic variables, your target population may be quite broad. In this case some sampling techniques such as random or stratified will probably be inaccessible to you for logistical reasons, and your choice of sampling method will be driven by accessibility of participants. If there are reasons to believe that the phenomenon is highly variable, you will probably narrow down your target population (for example, to students from your school) and will not claim that your results can be generalized beyond that. In this case the range of accessible sampling techniques is wider.

Once you have recruited the sample, make sure you identify its essential characteristics, such as age, gender distribution, education and whatever else is important to the aim of your investigation. For example, in a replication of Loftus and Palmer (1974) you might want to mention language proficiency because there are reasons to believe that participants will behave differently if the leading question is asked in a language that is not their mother tongue.

How do you explain the procedure of the experiment?

Explaining the procedure of your experiment involves the following elements: explaining the controlled variables, outlining the step-by-step procedure of the experiment, and explaining the choice of materials.

Explaining the controlled variables

Remember that you only need to control variables that can potentially be confounding for your experiment. This means that such variables first need to be identified. Think about what can affect the results of your study. Come up with a list of factors. For each of the factors explain how it can potentially interfere with the results. Range the factors and look at the most important ones. For each of them, decide which of the following you are going to do about them:

- do nothing and acknowledge they might be influencing your results
- use random allocation into groups and expect that these factors will influence all experimental conditions equally, so will not affect the comparison
- eliminate the confounding factors.

Outlining the step-by-step procedure of the experiment

The main characteristic of this section of your report is its replicability. Write it in a way that will allow an independent researcher to take the description and replicate the whole procedure step by step. In the process you will make references to materials. For example, if at a certain stage of the procedure standardized instructions were given to participants, you are expected to include the full text of the standardized instructions in an appendix and make a reference to it in the text. The same applies to the standard debriefing notes, informed consent form, stimulus materials that you used in the experiment, and so on.

Explaining the choice of materials

Some of the materials will need to be explained. For example, in a replication of Loftus and Palmer (1974), why did you use this particular video of a car accident? How many videos did you use and why? In a replication of Bransford and Johnson (1972), why did you choose to use this text? All the explanations you suggest should be supported by what works best to test your hypothesis and achieve the aim of your investigation.

Conducting the analysis

Let's go again to the *Psychology guide* and the assessment criteria against which your work will be assessed. For analysis (assessed against criterion C), the top markband for 5–6 marks states:

- Descriptive and inferential statistics are appropriately and accurately applied.
- The graph is correctly presented and addresses the hypothesis.
- The statistical findings are interpreted with regard to the data and linked to the hypothesis.

Let's suppose you followed our suggestion and conducted a simple experimental study with one IV, two conditions (or groups, depending on the experimental design) and one DV. Now you need to analyse the data and report the results. For this you will need to choose descriptive and inferential statistical tests. In this section we explain the rationale behind the choice of the tests that you are most likely to use in internal assessment, the calculations associated with the tests how to report them.

Descriptive statistics

As the name suggests, the purpose of descriptive statistics is to describe variables. You will need to determine the level of measurement of the DV, normality of its distribution and calculate a measure of central tendency and a measure of dispersion.

Levels of measurement

There are four levels of measurement, or types of variable.

Nominal-level variables cannot be quantified. They represent a set of labels that cannot be placed in either descending or ascending order. Examples of nominal variables are: car brands, zodiac signs, gender, nationality and music genre preferences. Some people prefer jazz and some listen to rock, but you cannot rank people by their music preferences from the lowest to the highest. Answers to yes/no questions and other **dichotomous variables** (that is, variables that can only take one of two values) are also usually treated as nominal data. Since nominal variables cannot be quantified; procedures such as addition, subtraction or multiplication cannot be applied to them. This affects the range of statistical tests that is possible.

Ordinal-level variables can be ranked from the lowest to the highest, but the intervals between ranks may be unequal. An example would be the results of a car race: competitors come first, second, third and so on, but the difference between the person who comes first and the one who finishes second is not necessarily the same as the difference between the second and the third finishers.

Interval-level variables are like ordinal variables, but the intervals are assumed to be equal. Importantly, these variables may have a zero value, but zero does not mean an "absence" of the parameter. An example would be an ordinary thermometer that measures temperature in Centigrade. You can say that the difference between 35°C and 37°C is the same as the difference between 31°C and 33°C. You also know that 0°C does not mean an "absence of temperature", it is just an arbitrary point on the scale. Due to these properties of interval variables, addition and subtraction can be used with them but multiplication and division cannot. For example, it is correct to say that 12°C + 5°C = 17°C, but it is incorrect to say that 20°C is twice as much as 10°C because if you move the zero point, that would no longer be true.

Finally, **ratio-level variables** are like interval variables, but the zero is fixed and meaningful—it means the absence of something. For example, the number of words correctly recalled from a list may be considered ratio data because if you recalled no words it simply means that you have not recalled anything! Annual salary, age and weight are all examples of ratio-level data. Building on our example above, if you measure temperature in Kelvin (where 0 degrees corresponds to minus 273°C and is known as the "absolute zero", the lowest temperature possible), then your data is measured on the ratio level. The full range of mathematical operations can be applied to ratio-level data including division and multiplication.

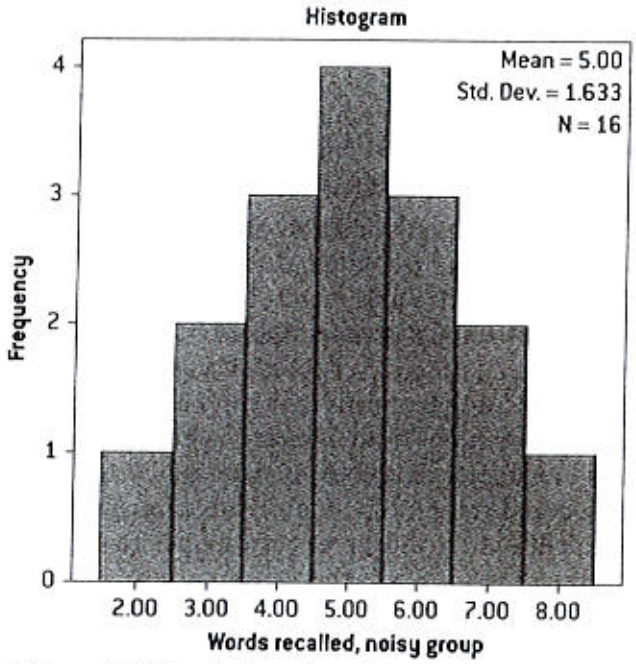
Normality of distribution of the dependent variable

Normality of distribution as a concept is only applicable to interval and ratio data. **Distribution** is the range of values of a variable according to their frequency. Let's use an example to clarify this.

Suppose you gave people a list of words under two conditions, quiet and noisy, and asked them to recall the words later, registering the number of words that they recalled correctly. These could be the numeric results of the investigation.

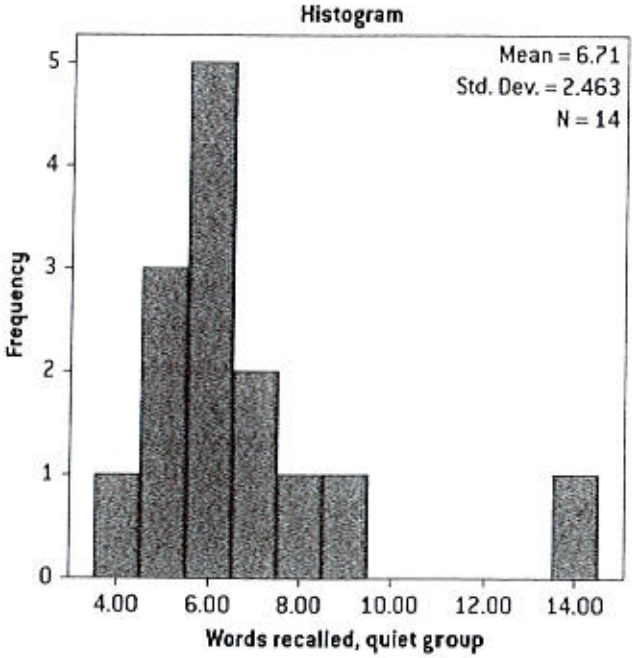
Experimental condition (noisy)		Control condition (quiet)	
Participant	Number of words recalled	Participant	Number of words recalled
1	5	1	14
2	4	2	5
3	6	3	9
4	3	4	4
5	4	5	6
6	5	6	5
7	5	7	6
8	6	8	5
9	4	9	6
10	6	10	7
11	7	11	8
12	5	12	6
13	3	13	7
14	7	14	6
15	8		
16	2		

If we plot the frequency distribution of the data from the experimental group, noisy, this is what we get:



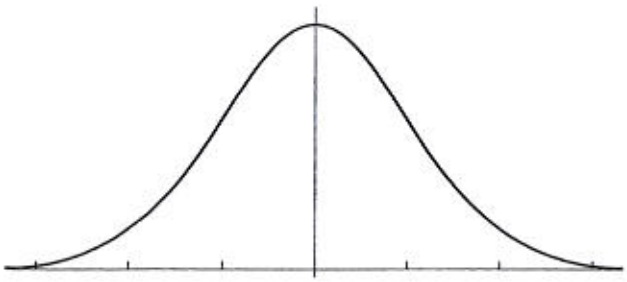
▲ Figure 9.1 Distribution of words recalled correctly, noisy condition

For the control group, quiet, we get the following distribution:



▲ Figure 9.2 Distribution of words recalled correctly, quiet condition

Normal distribution is a special type of distribution that visually looks like a bell. This is why it is sometimes called the "bell curve". Here is a visual representation of an ideal normal distribution:



▲ Figure 9.3 Normal distribution

Normal distribution occurs naturally in many situations. For example, IQ scores, height of people, age, blood pressure would all be normally distributed. Most people would be "average", and when the scores are more extreme, fewer people in the population would have these extreme scores.

In reality most data sets will deviate to some extent from the ideal normal distribution. How large is this deviation, however, and can we still assume that data are normally distributed even if minor deviations take place? Rigorous tests exist to compare the empirical distribution (the one you obtained in a study) to the ideal normal distribution and conclude whether the deviation is small enough to assume normality. However, these tests fall outside the scope of the internal assessment requirements; you just need to know the concept of normality of distribution and use the following general principles.

- Scan the raw data for outliers. An outlier is an observation that is extremely different from most or all other observations. For example, in the distribution plot for the control group in our fictitious data set, the participant who scored 14 (participant 1) may be called an outlier because the score is so different from all other scores. If severe outliers are present, the distribution is probably not normal, especially with small sample sizes.
- Optionally, use software to build the distribution of your data (current versions of MS Excel have this function, for example) and visually assess it. If its shape resembles a bell approximately (with average values being most frequent and extreme values less frequent), you may assume normality.

You must include your raw data table in an appendix in your report. It is not necessary to include the distribution plot, but your decision regarding normality of distribution will affect your choice of descriptive and inferential tests.

First, let's look at descriptive statistics: measures of central tendency and measures of dispersion. These measures help us summarize data to make sense of it.

Measures of central tendency

There are three measures of central tendency: the mean, the median and the mode.

The mean is simply the average of all scores. Add all the data points and divide them by the number of observations (participants in the group). The mean is the most common measure of central tendency, but it may be biased if there are some extreme outliers in the sample. The problem with outliers is that they can skew the mean so that it is no longer a meaningful indicator of central tendency, so you need to be careful in applying the mean to data sets with extreme outliers and maybe use the median or the mode instead.

The median is the "middle" of a sorted list of numbers. To find the median, place the numbers in value order. For example, suppose we have the following values:

3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 40, 56

There are fourteen values so the "middle" should be between the seventh and the eighth value: we just find the average. $\frac{(21 + 23)}{2} = 22$. The median of this data set is 22.

The median is used when there are strong deviations from normality of distribution making the mean misleading.

The mode is the value that appears most often in the data set. In our hypothetical experiment with the quiet and the noisy condition, the mode in both groups is 6. There are cases when a data set has more than one mode, for example, here is a binomial data set (the modes are 3 and 6):

1, 3, 3, 3, 4, 4, 6, 6, 6, 9

In an ideally normal distribution the mean, the median and the mode will coincide. As the distribution deviates from normality, especially if there are outliers, they may move apart. So when the level of measurement is interval or ratio,

your choice of the measure of central tendency should be guided by your assumptions about the normality of distribution in the data sets. When the level of measurement is ordinal, the mean cannot be used. When the level of measurement is nominal, you can only use the mode.

Measures of dispersion

To make sense of data, measures of central tendency are not enough because some data sets are more spread out around the centre than others. Two measures of dispersion are commonly used: the standard deviation and the semi-interquartile range.

Standard deviation is calculated using the following formula:

$$SD = \sqrt{\frac{\sum(x_i - \bar{X})^2}{N - 1}}$$

Let's decipher the formula step by step. First you calculate the mean of the dataset ($\bar{X}$), then for each individual value (x_i) you calculate its deviation from the mean ($x_i - \bar{X}$). After this you square the deviation, obtaining squared deviations ($(x_i - \bar{X})^2$). This eliminates the signs (positive and negative). You calculate the sum of all squared deviations and divide it by the number of observations (participants) minus one. This gives you the "average squared deviation". Finally, you calculate the square root of that value.

Standard deviation is an accurate measure of dispersion that takes into account all individual values and uses all information available in the data set, making it the preferred choice when possible. A low standard deviation indicates that the data points tend to be closely grouped around the mean whereas a high standard deviation indicates that the data points are further spread out around the mean.

However, since the formula uses the mean, the standard deviation cannot be used with nominal or ordinal data, or with data sets that severely violate normality of distribution. So, even if your level of measurement is interval or ratio, you might prefer other methods if there are outliers.

Semi-interquartile range, unlike standard deviation, does not assume normality of distribution and can be used with ordinal-level data.

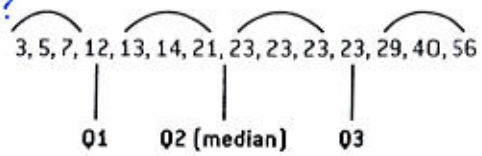
When you sort a list of numbers, you can find the so-called **quartiles**:

Q2 (the middle quartile) is the median

Q1 (the lower quartile) is the median of the numbers to the left of Q2

Q3 (the upper quartile) is the median of the numbers to the right of Q3

Taking our example from before:



▲ Figure 9.4 Finding quartiles

Q1 is the median of the dataset 3, 5, 7, 12, 13, 14, 21, that is, Q2 = 12

Q2 = the median = 22

Q3 is the median of the dataset 23, 23, 23, 23, 29, 40, 56, that is, Q3 = 23.

The semi-interquartile range is calculated simply as:

$$SIR = \frac{Q3 - Q1}{2}$$

In this example the semi-interquartile range is equal to $\frac{(23 - 12)}{2} = 5.5$

Similarly to the standard deviation, the higher the semi-interquartile range, the more the individual data points are spread out around the centre of the distribution.

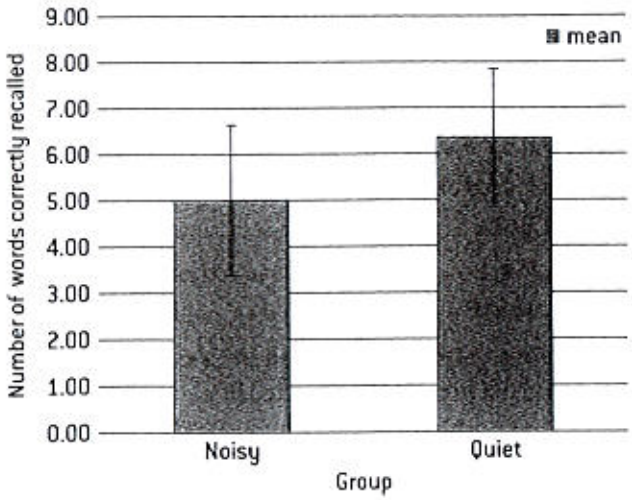
Graphing the results

Whatever measures of central tendency and dispersion you have chosen, you need to clearly show this information graphically. In the most typical scenario where you compare two conditions, a **bar chart with error bars** would be the most appropriate choice. Once you know your measures of central tendency and dispersion, it is easy to produce such charts in widely available software, for example, MS Excel.

Psychology in real life

If you use MS Excel to generate error bars, you will need to use the "custom values" option for your error bars and indicate the range of cells where your dispersion values are stored. You can find many tutorials online, such as this one: https://www.youtube.com/watch?v=s3_hcC0Gc50.

An example of a bar chart is shown below.



▲ Figure 9.5 Means and standard deviations of the number of words correctly recalled in the "noisy" and "quiet" conditions (note that error bars denote one standard deviation)

When making the bar chart, follow these guidelines.

- The bar chart should be clearly linked to the hypothesis. Do not include information that is not essential for the testing of your hypothesis.
- Do not forget to include clear and meaningful labels for both the axes. Avoid vague labels like "Group 1" and "Group 2".
- Mention the units of measurement for the DV if necessary. ↳ Pol forget this all the time
- Include a clear title that is self-explanatory.
- The y-axis should start with the lowest possible value, not with the lowest value observed in your data. Note how the y-axis in our example starts with a zero.

Inferential statistics

While descriptive statistics are used to describe the DV, inferential statistics are used to test hypotheses about its relationship with the IV.

Your choice of the inferential test will depend on:

- the experimental design
- the level of measurement
- your assumptions regarding normality of distribution of the DV.

Below you will find an overview of some of the appropriate inferential tests. Note that all these tests are only suitable for situations when the IV has two levels (that is, you are comparing either two groups or two conditions).

★★ Key Chart! Do not skip past this ★★

		Experimental design			
		Independent measures		Repeated measures (or matched pairs)	
Level of measurement	Nominal	Chi-squared (χ^2) test		McNemar's test (for dichotomous data)	
	Ordinal	Mann–Whitney U test		Wilcoxon signed-rank test	
	Interval / Ratio	Mann–Whitney U test	Unrelated t-test	Wilcoxon signed-rank test	Related t-test

Each inferential test has some assumptions and can only be used if these assumptions are met.

First, some tests assume normality of distribution. Such tests are called **parametric** because they use "parameters" (mean and standard deviation) in their formula, and, as you know, neither the mean nor the standard deviation are meaningful under severe violations of normality of distribution. The related t-test and unrelated t-test are parametric. All the other tests in the table above are **non-parametric**: they do not assume normality of distribution because they do not use the mean or the standard deviation in the calculations.

So you probably won't use them; even if you do have normally distr. data there's still more conditions that need to be met to do one.

Second, tests assume a certain level of measurement. As you know, the level of measurement determines what mathematical operations can be performed with a variable, which is why there are restrictions. For example, the Mann–Whitney U test assumes at least the ordinal level of measurement. It means the test cannot be used with data below that level. It can be used with interval and ratio data but such data will be first reduced to the ordinal level of measurement. ★

↳ see?

Third, inferential tests are specific to the experimental design. For example, there are two types of t-test, one for unrelated samples (independent measures design) and one for related samples (repeated measures or matched pairs design). You need to take care in applying the appropriate inferential test.

Bad example - in order replace w/ Mann-Whitney U & Wilcoxon

Exam tip

It is helpful to know the calculation procedure and to be able to compute your chosen test manually, and you are encouraged to learn how to do that. However, manual calculations are not required for internal assessment purposes, you can use software. Below we will focus on the four most commonly used tests (unrelated t-test, related t-test, Mann–Whitney U test and Wilcoxon signed-rank test), looking at the calculation process and reporting the results of these tests. Although you can use software, understanding of the calculations will help you greatly in terms of knowing which option to choose and how to interpret the output.

You will do this

Only if you want to verify → kids who just do it on a calcul. get points off all the time bc they don't know how to interpret the results.

**

Should I choose a parametric or a non-parametric test?

On the one hand, parametric tests have greater power, that is, they are more capable of detecting existing differences. Non-parametric tests reduce the level of measurement to ordinal, ignoring the information about the intervals between data points, and therefore lose some power. It is more likely that they will not detect an existing difference. The difference in power is not great: under normal distribution non-parametric tests will detect the existing difference as well as parametric tests in 95% of cases. However, parametric tests can still be considered to be more sensitive.

Oh well.

That's ok. You can have a perfectly valid exp. if the data is not significant

On the other hand, parametric tests depend on a number of assumptions, most importantly, normality of distribution. With small sample sizes there is no rigorous way to assess normality of distribution, so we can only rely on the visual analysis of the data for obvious outliers. When we are not too sure that the distribution is normal, it might be safer to prefer non-parametric methods. — H 3.

Unrelated t-test

The unrelated t-test is used to test the difference between two independent samples. It is also known as independent t-test, independent sample t-test and Student's t-test for independent samples.

The unrelated t-test relies on the following assumptions.

- The level of measurement is either interval or ratio.
- You are using an independent measures design with two groups of participants.
- The DV should be approximately normally distributed for each group of the IV. We say "approximately" because the t-test has been shown to be quite robust to moderate violations of normality. This means that it is still valid even if data somewhat deviates from normality. However, severe deviations may compromise the validity of the test. If there are doubts, the safest option is to choose a non-parametric test instead. When looking at outliers and normality of distribution, remember to look separately at the two groups.
- There is homogeneity of variances in the two groups. This means that the standard deviations in the two groups should be approximately equal or at least comparable. However, there is a modification of the unrelated t-test that can handle data when homogeneity of variances is not met. As a general rule, we suggest that you always use this modification, especially when the sample size is small.

The formula of the unrelated t-test is:

$$t = \frac{M_1 - M_2}{\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}}$$

where M, SD and n are the means, standard deviations and sample sizes in the two compared groups respectively. Note that equal variances are not assumed in this formula, so it is robust to the violation of the fourth assumption outlined above.

Let's go back to our example, but change the value for the first participant in the control group to get rid of the outlier.

Experimental condition (noisy)		Control condition (quiet)	
Participant	**Number of words recalled**	**Participant**	**Number of words recalled**
1	5	**1**	9
2	4	**2**	5
3	6	**3**	9
4	3	**4**	4
5	4	**5**	6
6	5	**6**	5
7	5	**7**	6
8	6	**8**	5
9	4	**9**	6
10	6	**10**	7
11	7	**11**	8
12	5	**12**	6
13	3	**13**	7
14	7	**14**	6
15	8		
16	2		
Mean	**5.0**		**6.36**
SD	**1.63**		**1.50**

Plugging the values into the formula, we get:

$$t = \frac{5 - 6.36}{\sqrt{\frac{1.63^2}{16} + \frac{1.50^2}{14}}} = -2.38$$

So the test statistic for the unrelated t-test is -2.38. To find out its statistical significance, we compare this value to the table of critical values easily found on the internet (just search for "t-test critical values table").

Here is a part of that table:

Critical values for the t-test:							
	one-tailed	**.05**	**.025**	**.01**	**.005**	**.001**	**.0005**
df	**two-tailed**	**.1**	**.05**	**.02**	**.01**	**.002**	**.001**
1		6.314	12.706	31.821	63.657	318.31	636.62
2		2.920	4.303	6.965	9.925	22.327	31.598
3		2.353	3.182	4.541	5.841	10.214	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
13		1.771	2.160	2.650	3.012	3.852	4.221
26		1.706	2.056	2.479	2.779	3.435	3.707
27		1.703	2.052	2.473	2.771	3.421	3.690
28		1.701	2.048	2.467	2.763	3.408	3.674
29		1.699	2.045	2.462	2.756	3.396	3.659
120		1.658	1.980	2.358	2.6117	3.160	3.373

First, we choose whether the test is one-tailed or two-tailed. It depends on your hypothesis. If it predicted the direction of the relationship between variables (for example, "A will be higher than B"), then the test is directional. Supposing we had predicted that recall in the "quiet" group would

be higher, we would need to look at levels of significance for a one-tailed test.

The next piece of information we need is degrees of freedom (df). For the unrelated t-test degrees of freedom are simply N (the total number of participants) minus 2. So for our example df = 30 – 2 = 28.

The sign of the t statistic can be ignored for now, because it only shows which of the groups has a higher mean. In line with our hypothesis, we see that the mean for the "quiet" group (6.36) is higher than the mean for the "noisy" group (5.0).

The cut-off level of significance, as you know, is $p < 0.05$. According to the table, the critical value for a one-tailed t test at $p = 0.05$ and 28 degrees of freedom is 1.701. Our value (2.38) is greater than that, so we need to conclude that the result is statistically significant. We should reject the null hypothesis and accept the experimental hypothesis.

Results of the test will be reported like this: recall in the "quiet" condition ($M = 6.36$, $SD = 1.50$) was significantly better than recall in the "noisy" condition ($M = 5.0$, $SD = 1.63$); $t(28) = -2.38$, $p < 0.05$, one-tailed.

Should I choose the unrelated t-test or the Mann–Whitney U test?

Reviewing all considerations in the context of a typical internal assessment project in IB psychology, our recommendation would be to choose the unrelated t-test in the following situations.

- Your sample size is sufficient, for example, no fewer than 15 participants in each group so that you can visually detect outliers.
- Group sizes are approximately equal, that is, the ratio of the largest to the smallest group size is not larger than 1.5.
- There are no apparent outliers in any of the two groups and the distributions do not seem to deviate severely from normality.

If one or more of these criteria are not met, we suggest it would be safer to use the non-parametric test instead.

Related t-test

The related t-test is used to compare means in two related (dependent) samples.

It relies on the following assumptions.

- The level of measurement is either interval or ratio.
- You are using either a repeated measures or matched pairs design.
- The distribution of the differences in the DV between the two related groups should be approximately normal. We will explain this in the example below.

The formula for the test is:

$$t = \frac{M_{diff}}{\frac{SD_{diff}}{\sqrt{n}}}$$

Where M_{diff} is the mean of the differences, SD_{diff} is the standard deviation of the differences and n is the sample size.

Suppose that you have conducted the same "quiet versus noisy" experiment, but this time you have used a repeated measures design (the same participants were tested twice under different conditions). You make a raw data table and for each participant calculate the difference score (quiet minus noisy).

Participant	Control condition (quiet)	Experimental condition (noisy)	Difference (quiet minus noisy)
1	9	5	4
2	5	4	1
3	9	6	3
4	4	3	1
5	6	4	2

Participant	Control condition (quiet)	Experimental condition (noisy)	Difference (quiet minus noisy)
6	5	5	0
7	6	5	1
8	5	6	−1
9	6	4	2
10	7	6	1
11	8	7	1
12	6	5	1
13	7	3	4
14	6	7	−1
Mean	**6.36**	**5.0**	**1.36**
SD	**1.50**	**1.30**	**1.55**

The third assumption of the related t-test tells you that the distribution of difference scores (the values in the last column of the table) should be approximately normal. There are no obvious outliers, so let us assume it is so.

Plugging the values into the formula, we get:

$$t = \frac{1.36}{\frac{1.55}{\sqrt{14}}} = 3.28$$

The number of degrees of freedom for this test is n−1. The critical values table for the related and the unrelated t-test is the same, so for our example the critical value for a one-tailed test at the $p = 0.05$ level of significance for 13 degrees of freedom is 1.771. Our value is greater than that, so we can report: there was a significant difference in the scores for the quiet condition ($M = 6.36$, $SD = 2.25$) and the noisy condition ($M = 5.0$, $SD = 1.69$), with the quiet condition resulting in better recall; $t(13) = 3.28$, $p < 0.05$, one-tailed.

Should I choose the related t-test or the Wilcoxon signed-rank test?

Reviewing all considerations in the context of a typical internal assessment project in IB psychology, our recommendation would be to choose the related t-test in the following situations.

- Your sample size is sufficient, for example, no less than 15 participants so that you can visually detect outliers.
- There are no apparent outliers in the difference scores and the distribution of the difference scores does not seem to deviate severely from normality.

If one or more of these criteria are not met, we suggest it would be safer to use the non-parametric test instead.

Mann–Whitney U test

The Mann–Whitney U test is a non-parametric equivalent of the unrelated t-test.

It has the following assumptions.

- The level of measurement is at least ordinal.
- You have used an independent measures design.

The Mann–Whitney U test does not rely on any assumptions regarding the distribution of the DV, making it non-parametric, so the test uses neither the mean nor the standard deviation in the formula. Instead it reduces the level of measurement to ordinal and uses ranks.

Importantly, the test can only be interpreted as comparing two medians if some additional assumptions are met. Without these assumptions the test is still valid, but should be interpreted as a test of difference in mean ranks. That is exactly what we will do.

Let's take the same data as we used in the example for the unrelated t-test:

Experimental condition (noisy)		Control condition (quiet)	
Participant	Number of words recalled	Participant	Number of words recalled
1	5	1	9
2	4	2	5
3	6	3	9
4	3	4	4
5	4	5	6
6	5	6	5
7	5	7	6
8	6	8	5
9	4	9	6
10	6	10	7
11	7	11	8
12	5	12	6
13	3	13	7
14	7	14	6
15	8		
16	2		

First we need to rank the data, so as to reduce the level of measurement to ordinal and avoid the need to use the mean and the standard deviation. There are four steps.

1. Grouping—we order all values, regardless of the group, from the smallest to the largest, placing all "ties" in the same lines.
2. Ordering—we assign these values ranks, from the lowest to the highest, ignoring the ties.
3. Ranking—we replace the ranks for the ties with their respective means, so that equal values get the same ranks.
4. We calculate the mean rank and the sum of ranks in each group.

Step 1 (grouping)		Step 2 (ordering)		Step 3 (ranking)	
Control condition (quiet)	Experimental condition (noisy)	Control condition (quiet)	Experimental condition (noisy)	Control condition (quiet)	Experimental condition (noisy)
	2		1		1
	3		2		2.5
	3		3		2.5
4	4	4	5	5.5	5.5
	4		6		5.5
	4		7		5.5
5	5	8	9	11	11
5	5	10	11	11	11
5	5	12	13	11	11
	5		14		11
6	6	15	16	18.5	18.5
6	6	17	18	18.5	18.5
6	6	19	20	18.5	18.5
6		21		18.5	
6		22		18.5	
7	7	23	24	24.5	24.5
7	7	25	26	24.5	24.5
8	8	27	28	27.5	27.5
9		29		29.5	
9		30		29.5	
			Mean rank	19.04	12.41
			Sum of ranks	266.5	198.5

Now we plug these values into the formulas:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_2$$

where R_1 and R_2 are sums of ranks for the two groups respectively, and n_1 and n_2 are the group sizes.

For our example we get:

$$U_1 = 14 \times 16 + \frac{14 \times (14 + 1)}{2} - 266.5 = 62.5$$

$$U_2 = 14 \times 16 + \frac{14 \times (14 + 1)}{2} - 198.5 = 130.5$$

As the resulting test statistic for Mann–Whitney, we simply choose the smallest of U_1 and U_2, so for our example the test statistic U = 62.5.

To find the critical value, you just need to know the group sizes. The table critical value for a one-tailed U test at p = 0.05 for n_1 = 14 and n_2 = 16 is 71. To be significant, the empirical value must be equal to or smaller than the table value, so for our example, since 62.5 < 71, we may conclude that the result is statistically significant and report: a Mann–Whitney U test indicated that recall in the quiet condition (mean rank = 19.04) was significantly higher than recall in the noisy condition (mean rank = 12.41); U(14,16) = 62.5, p < 0.05, one-tailed.

Wilcoxon signed-rank test

The Wilcoxon signed-rank test is a non-parametric equivalent of the related t-test.

It has the following assumptions.

- The level of measurement is at least ordinal.
- You are using a repeated measures or matched pairs design.

Participant	Control condition (quiet)	Experimental condition (noisy)	Difference (quiet minus noisy)
1	9	5	4
2	5	4	1
3	9	6	3
4	4	3	1
5	6	4	2
6	5	5	0
7	6	5	1
8	5	6	−1
9	6	4	2
10	7	6	1
11	8	7	1
12	6	5	1
13	7	3	4
14	6	7	−1

Just like the Mann–Whitney U test, the Wilcoxon signed-rank test reduces the level of measurement to ordinal, and just like the related t-test, it deals with differences between pairs of values.

Let's take the same example we used for the related t-test (table at the top of the page).

After calculating the difference scores (the last column of the table), we need to do some transformations with the obtained differences. There are four steps.

1. We order them by their absolute value (ignoring the sign).
2. We assign ranks while still ignoring the signs. At this point we eliminate data points where the difference is zero, thus "reducing" the sample size. Remember to observe the ties.
3. We return the signs to the ranks.
4. We calculate the sum of positive ranks (W+) and the sum of negative ranks (W−).

Step 1	Step 2	Step 3
Difference (quiet minus noisy), ordered (sign ignored)	Ranked difference (sign ignored)	Sign returned
0	---	---
1	4.5	+ 4.5
1	4.5	+ 4.5
1	4.5	+ 4.5
−1	4.5	− 4.5
1	4.5	+ 4.5
1	4.5	+ 4.5
1	4.5	+ 4.5
−1	4.5	− 4.5
2	9.5	+ 9.5
2	9.5	+ 9.5
3	11	+ 11
4	12.5	+ 12.5
4	12.5	+ 12.5
W+ (the sum of positive ranks)		82
W− (the sum of negative ranks)		9

W+ in this example is larger, and this tells you that recall in the quiet condition was generally better than recall in the noisy condition (because we subtracted noisy from quiet). However, we still need to test this difference for statistical significance.

The test statistic for Wilcoxon signed-rank test is denoted by the letter T and is simply the smallest of W+ and W–. In this case T = 9.

The table of critical values for Wilcoxon signed-rank test gives 21, 12 and 4 as the values for a one-tailed test with n = 13 (remember we "reduced" the sample by one!) and at p = 0.05, p = 0.01 and p = 0.001 respectively. The result is significant if the obtained value is equal to or smaller than the critical value. In our case 9 is smaller than 12 so we can conclude that the result is significant at p < 0 .01. We should reject the null hypothesis and accept the experimental hypothesis. We can report: a Wilcoxon signed-rank test indicated that recall in the quiet condition was significantly better than recall in the noisy condition: T = 11, p < 0.01, one-tailed.

Note that if you do calculations by hand and compare the obtained value to the value of critical tables, you will be using the T statistic because it is easier and tables of critical values are made for it. However, if you use a statistical package it may report the Z statistic instead. Do not be confused by that; both T and Z are statistics related to the Wilcoxon test. The software will calculate the level of significance automatically, so all you need to do is to report the Z value (and the corresponding level of significance) instead of T.

Useful resources

You will find the following resources helpful.

Tables of critical values:

- t-tests: https://tinyurl.com/5wtxc9h
- Mann–Whitney U test: https://tinyurl.com/y8j2xrhm
- Wilcoxon signed-rank test: https://tinyurl.com/yctlm2av

Online statistical software:

- GraphPad allows you to calculate descriptive statistics and t-tests; it also has a feature to run tests to detect outliers in your data: http://graphpad.com/quickcalcs/contMenu/
- VassarStats has a variety of statistical tests that can be run online: http://vassarstats.net

Writing the evaluation

Let's go back again to the *Psychology guide* and the assessment criteria against which your work will be assessed. For evaluation (assessed against criterion D) the top markband for 5–6 marks states:

- The findings of the student's investigation are discussed with reference to the background theory or model.
- Strengths and limitations of the design, sample and procedure are stated and explained and relevant to the investigation.
- Modifications are explicitly linked to the limitations of the student's investigation and fully justified.

The final step of your investigation and the last section of your report is evaluation. This should include three major elements: linking the findings to the background theory or model, analysing strengths and limitations, and suggesting modifications.

When you link the findings to the background theory or model, the main question to answer is whether or not your findings support the background theory and why. Aspects of this discussion may include the comparison between your findings and those of the original experiment. You can also reiterate the modifications you have made to the original procedure and discuss how this might have influenced the discrepancy (if any). But most importantly, this is where you switch from the operationalized language of hypothesis testing to the language of constructs again. If you were replicating Loftus and Palmer, you should discuss what your findings mean in the context of the theory of reconstructive memory: do your findings support the idea that post-event information may influence memory of the event? If you were replicating Bransford and Johnson, you should discuss your findings in the context of schema and prior contextual information influencing comprehension and recall of verbal material.

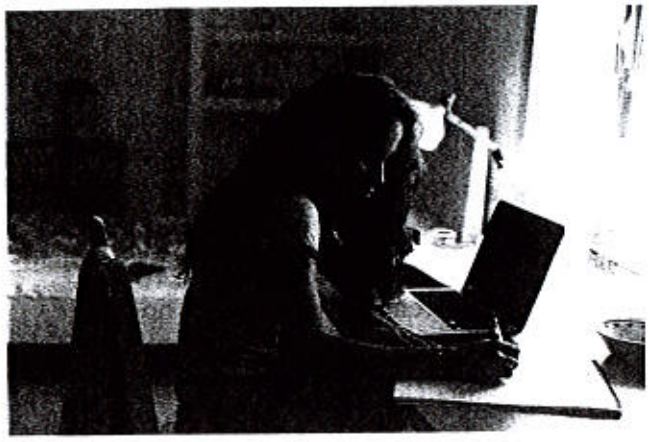
When you analyse the strengths and limitations of your study, keep in mind the following considerations.

- Your analysis should be comprehensive. Take into account construct validity, internal validity and external validity of your experiment. Consider how the constructs were operationalized, what variables were controlled, how the sample was selected, what the target population was, and to what extent various forms of bias might have taken place.
- Your analysis should be directly relevant to your investigation. Avoid general statements that are applicable to almost any experimental study. A typical mistake that students make is to say that since the study was a laboratory experiment, it lacked ecological validity, and to say nothing else. This might be a valid point, but why is this an issue particularly in your experiment as compared to other research?
- Talk about both strengths and limitations.
- Draw a line between methodological limitations of your experiment and limitations in the quality of its implementation. For example, suppose in a replication of Bransford and Johnson (1972) you decided to use a filler task between listening to the text and reproducing it. This decision was justified: you said that the filler task is necessary to avoid mechanical repetition and only measure information retained in the long-term memory. Later you said that using a filler task is a limitation of the experimental design because it extends the duration of the study and makes possible such threats to internal validity as history effect. This is reasonable: you made a choice, justified it, but recognized the limitations of this choice in terms of potential threats to internal validity. This counts as a methodological limitation of the experiment. In contrast, suppose participants in your Bransford and Johnson replication were tested as a group. They listened to the text and then engaged in the filler activity. While doing that, they occasionally talked to each other or distracted each other. You cannot say that this is a methodological limitation of your experiment: it is just bad implementation!

Methodological and other limitations of the study you have identified will be closely linked to suggested modifications. Continuing our example

Best mods suggest new ways to run the exp. to test the same basic variables

with the filler task activity in a replication of Bransford and Johnson, you might suggest, for example, that future studies vary the duration of the filler activity and assess its impact on the results. All potential confounding variables cannot be possibly eliminated in a single research study, so it is perfectly fine to suggest modifications for future research endeavours. However, avoid generic suggestions such as increasing the sample size or conducting the study in more naturalistic settings. The modifications you suggest should target your hypothesis and be potentially informative in the context of the background theory or model.



Quieter, distraction-free environment

↳ These types of "mods" get you no points - they lack rigor.

References and appendices

References - USE: APA! (Yes, I'm yelling)

Remember that a failure to acknowledge the ideas of others is a breach of academic honesty. It is important to reference properly to acknowledge all ideas that are not yours in an academically acceptable manner. We are not going to outline the basic principles of academic honesty here because it is expected that you will already be familiar with them, but here are some guidelines for formatting your in-text citations and list of references, to avoid some common mistakes.

- For every in-text citation in the report there must be a corresponding entry in the list of references. Similarly, you should not include in the list of references anything that does not appear as a citation in the text.
- Your references must include all necessary information required by the citation style. For example, for a journal article you need to include the surnames and initials of the authors, year of publication, full name of the article and the journal, the number of journal or issue, and page numbers. It is good practice to make full references as you are doing the theoretical research so that you will not have to look for all these details again later.
- Your references must consistently follow one citation style, for example APA, MLA or Chicago. Strictly stick to the rules of your citation style and follow it throughout your report.

One useful resource is the Purdue Online Writing Lab. It provides guidance for formatting your papers in accordance with a variety of styles. Just choose your preferred citation style and explore the links: https://owl.english.purdue.edu/owl/section/2.

Appendices

The appendices should include:

- any standardized instructions used in the experiment
- briefing and debriefing notes
- a copy of the consent form (blank)
- standardized stimulus materials used in the experiment (blank - i.e. surveys should be incl. but w/o participant work on them)
- a raw data table
- calculations for both descriptive and inferential statistics—it is not a requirement to do calculations manually, but if you are using software you must include a print-out of the output with a clear indication of the options you have chosen (for example, one-tailed or two-tailed) and the values you are using as your results
- any other materials necessary for your investigation.
  - ↳ YouTube
  - ↳ PPT
  - ↳ Images
  - ↳ Words list etc.