(for example, a laboratory experiment), but you want to believe that people will behave the same way in their natural setting in daily life too. In any case, generalizability is an important aspect in the interpretation of findings. Again, the ways in which quantitative and qualitative research studies approach generalizability of findings is distinctly different.

The table below gives you an overview of the main concepts used to characterize sampling, generalizability, credibility and bias in experimental, correlational and qualitative research. As you read on, you will understand these concepts better. Refer to this table from time to time so that you place them clearly in the general framework.

Overview table:
Sampling, generalizability, credibility and bias in qualitative and quantitative research

	Quantitative research		Qualitative research
Overarching concepts	Experimental studies	Correlational studies	
Sampling	Random Stratified Self-selected Opportunity	Same	Quota sampling Purposive sampling Theoretical sampling Snowball sampling Convenience sampling
Generalizability	External validity: – Population validity – Ecological validity Construct validity	Population validity Construct validity	Sample-to-population generalization Case-to-case generalization Theoretical generalization
Credibility	Internal validity: to what extent is the DV influenced by the IV and not some other variable? Controlling confounding variables: eliminating or keeping constant in all conditions	No special term used: "validity" and "credibility" can be used interchangeably Credibility is high if no bias occurred	Credibility = trustworthiness. To what extent do the findings reflect the reality? Triangulation Establishing a rapport Iterative questioning Reflexivity Credibility checks Thick descriptions
Bias	Threats to internal validity: – Selection – History – Maturation – Testing effect – Instrumentation – Regression to the mean – Experimental mortality – Experimenter bias – Demand characteristics	On the level of measurement of variables: depends on the method of measurement On the level of interpretation of findings: – Curvilinear relationships – The third variable problem – Spurious correlations	Participant bias: – Acquiescence – Social desirability – Dominant respondent – Sensitivity Researcher bias: – Confirmation bias – Leading questions bias – Question order bias – Sampling bias – Biased reporting

▲ Table 1.2

Quantitative research: the experiment

Inquiry questions

- Why do experiments allow cause-and-effect inferences?
- How can bias in experimental research be prevented?
- How can findings from a small group of people be generalized to an entire population?
- How can experiments be designed?

What you will learn in this section

- Confounding variables
- Sampling in the experiment
  - Representativeness
  - Random sampling
  - Stratified sampling
  - Opportunity sampling
  - Self-selected sampling
- Experimental designs
  - Independent measures design
  - Matched pairs design; matching variable
  - Repeated measures design; order effects; counterbalancing
- Credibility and generalizability in the experiment: types of validity
  - Construct validity
  - Internal validity
  - External validity: population and ecological
- Bias in experimental research: threats to internal validity
  - Selection
  - History
  - Maturation
  - Testing effect
  - Instrumentation
  - Regression to the mean
  - Mortality
  - Demand characteristics
  - Experimenter bias
- Quasi-experiments versus true experiments
- Natural experiments and field experiments

Confounding variables

As we mentioned, the experiment is the only method that allows researchers to make cause-and-effect inferences. This is achieved by defining the independent variable (IV) and the dependent variable (DV), manipulating the IV and observing how the DV changes in response to this manipulation.

Psychological reality is very complex and the trick is to isolate the IV so that when you manipulate it, nothing else changes. Imagine, for example, that you manipulate X and observe the resulting changes in Y. However, every time you manipulate X, you also unintentionally change Z. In reality it is Z that causes a change in Y, but you incorrectly conclude that X (your IV) is the cause of Y, thus incorrectly confirming your hypothesis. If this sounds too abstract, think about the following example: X is sleep deprivation (which you manipulate by waking up one group of participants every 15 minutes when they sleep, while the control group sleeps normally) and Y is memory performance (which you measure by a simple memory test in the morning). Without realizing that this might be an important factor, you let the control group sleep at home while the experimental group sleeps in a laboratory being supervised by an experimenter. So there's another variable, variable Z: stress caused by the unfamiliar environment. It could be the case that in this experiment it was the unfamiliar environment (Z) that caused a reduction in memory performance (Y), rather than sleep deprivation (X).

Variables that can potentially distort the relationship between the IV and the DV (like Z in the example above) are called **confounding variables**. They contribute to bias. These variables need to be controlled, either by eliminating them or keeping them constant in all groups of participants so that they do not affect the comparison.

Discussion

How could the researchers have controlled the confounding variable in this example?

Sampling in the experiment

Being a truly nomothetic method, the experiment aims at discovering universal laws of behaviour applicable to large groups of people across a variety of situations. This makes relevant the distinction between the **sample** and the **target population**. The target population is the group of people to which the findings of the study are expected to be generalized. The sample is the group of people taking part in the experiment itself. How can we ensure that whatever results are obtained in the sample can be generalized to the target population? We do this through **representativeness**—the key property of a sample. A sample is said to be representative of the target population if it reflects all its essential characteristics.

Exercise

Imagine you are investigating the influence of praise on the school performance of teenagers. For this experiment you need to have a sample of participants that you would split into two groups (experimental and control). In the experimental group the teacher is instructed to praise every student three times a week while in the control group the teacher is told to only praise the students once every week. At the end of the research period performance grades in the two groups are compared.

Suppose that the participants in this experiment are high school students from one of the schools in your city. Will you be able to generalize the findings to the target population, that is, teenagers in general? This depends on how representative your sample is. For this you need to take into account your target population and the aim of the research.

- The aim of the research links to the **participant characteristics** that are essential. Whatever can theoretically influence the relationship between the IV and the DV is essential. For example, cultural background may be essential for how a teenager reacts to praise (depending on that teenager's cultural attitudes to adults, teachers and authority in general). Socio-economic background may be important as well: theoretically there may be a connection between the socio-economic status of a teenager's family and their value of education. The type of school is another potentially important factor: in top schools where students pursue quality education and prestigious college placements teachers' praise may be a point of pride, whereas in public schools in criminal neighbourhoods it may lead to bullying from classmates.
- If the sample is representative, it must reflect the essential characteristics of the target population. Is the sample of teenagers from one school in our example sufficient to reflect all these characteristics? No, because it does not represent the variation of cultural backgrounds, socio-economic backgrounds and types of schools found in the population.
- If the sample is not representative of the essential characteristics of the target population, there are two ways to fix it: either keep sampling or narrow down the target population and do not claim that the research findings are more generalizable than they really are.

Given the aim of the study, how would you increase representativeness of your sample?

There is no quantitative way to establish representativeness of a sample and it is always the expert decision of a researcher to say whether a particular characteristic is essential or not. This is done on the basis of prior knowledge from published theories and research studies. In any case the choice of the target population needs to be well justified and explicitly explained.

Several **sampling techniques** can be used in an experiment. The choice depends on the aim of the research, available resources and the nature of the target population.

- **Random sampling**. This is the ideal approach to make the sample representative. In random sampling every member of the target population has an equal chance of becoming part of the sample. With a sufficient sample size this means that you take into account all possible essential characteristics of the target population, even the ones you never suspected to play a role. Arguably, a random sample of sufficient size is a good representation of a population, making the results easily generalizable. However, random sampling is not always possible for practical reasons. If your target population is large, for example, all teenagers in the world, it is impossible to ensure that each member of this population gets an equal chance to enter your sample. Being based in Europe, you cannot just create a list of all teenagers in the world, randomly select a sample and then call Lynn from Fiji to come and join your experiment. In this case you either believe that cross-cultural differences are not essential (for your hypothesis) or narrow down your target population. On the other hand, if your target population is students from your school, it is perfectly possible to create the full list of students and select your participants randomly from this list. An example of random sampling strategy is a pre-election telephone survey where participants are selected randomly from the telephone book (or a random selection of Facebook profiles). Even in this case, though, you have to admit that the target population is not all the citizens of a particular country; it is all the citizens of the country who own a telephone (or have a Facebook profile).
- **Stratified sampling**. This approach is more theory-driven. First you decide the essential characteristics the sample has to reflect. Then you study the distribution of these characteristics in the target population (for this you may use statistical data available from various agencies). Then you recruit your participants in a way that keeps the same proportions in the sample as is observed in the population. For example, imagine that your target population is all the students in your school. The characteristics you decide are important for the aim of the study are age (primary school, middle school, high school) and grade point average—GPA (low, average, high). You study school records and find out the distribution of students across these categories:

	Low GPA	Average GPA	High GPA	Total
Primary school	0%	10%	10%	20%
Middle school	5%	30%	15%	50%
High school	5%	20%	5%	30%
Total	10%	60%	30%	100%

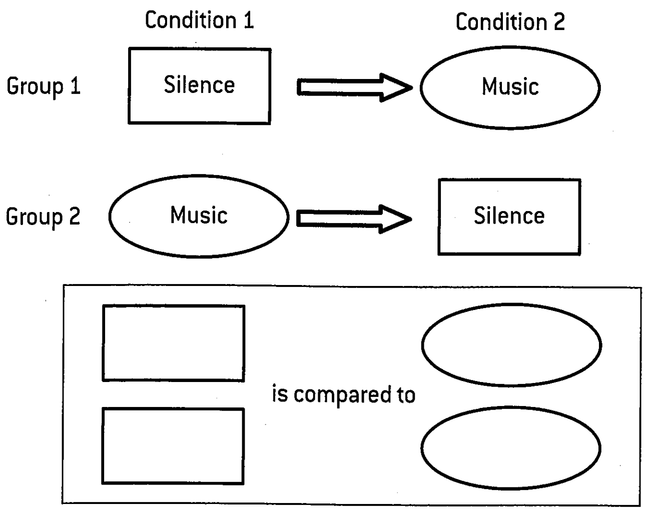
▲ Table 1.3

For a stratified sample you need to ensure that your sample has the same proportions. For every cell of this table you can either sample randomly or use other approaches (see below). In any case, what makes stratified sampling special is that it is theory-driven and it ensures that theory-defined essential characteristics of the population are fairly and equally represented in the sample. This may be the ideal choice when you are certain about essential participant characteristics and when available sample sizes are not large.

- **Convenience (opportunity) sampling**. For this technique you recruit participants that are more easily available. For example, university students are a very popular choice because researchers are usually also university professors so it is easy for them to find samples there. Jokingly, psychology has been sometimes referred to as a study of "US college freshmen and white rats". There could be several reasons for choosing convenience sampling. First, it is the technique of choice when financial resources and time are limited. Second, there may be different depending on which condition comes first (for example, silence then classical music or classical music then silence). Order effects may appear due to various reasons, such as the following.

- Practise: participants practise, improve their on-task concentration and become more comfortable with the experimental task during the first trial. Their performance in the second trial increases.
- Fatigue: participants get tired during the first trial, and their concentration decreases. Their performance in the second trial decreases.

To overcome order effects researchers use **counterbalancing**. Counterbalancing involves using other groups of participants where the order of the conditions is reversed. For our example, two groups could be used: one given the sequence "silence then music" and one given the sequence "music then silence". It is important to note that comparison will still be made between conditions, not between groups. Data from group 1 condition 1 will be collated with data from group 2 condition 2, and vice versa. These two collated data sets will be compared.



▲ Figure 1.4 Counterbalancing

An advantage of repeated measures designs is that people are essentially compared to themselves, which overcomes the influence of **participant variability** (differences between the groups before the experiment starts). It makes the comparison more reliable. Another advantage following from this is that smaller sample sizes are required.

Credibility and generalizability in the experiment: types of validity

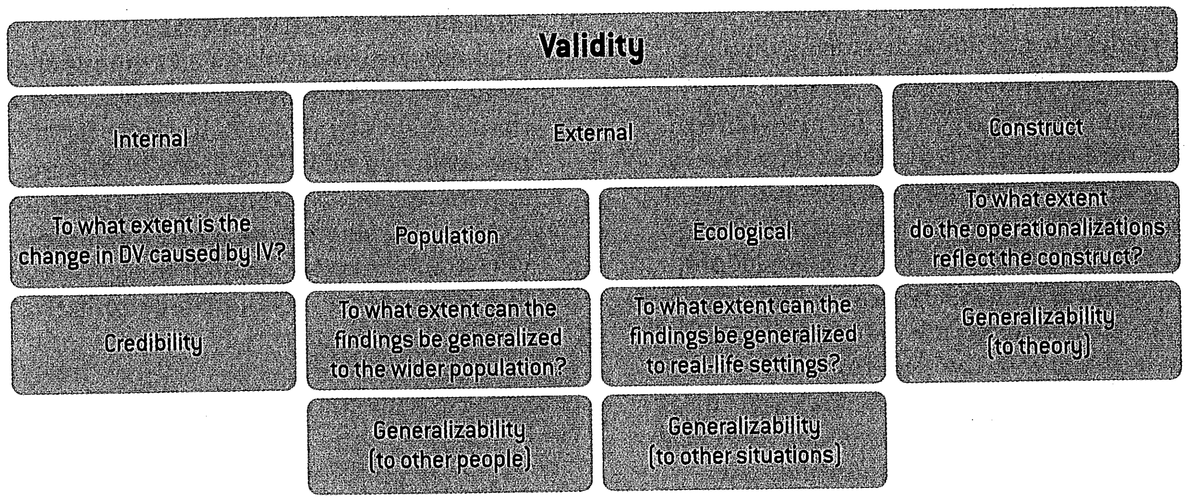
As you have seen, credibility and generalizability are overarching terms that are used to characterize the quality of research studies. When it comes to experiments specifically, these terms are very rarely used. Instead the quality of experiments is characterized by their construct, internal and external validity.

Construct validity characterizes the quality of operationalizations. As you know, the phenomenon under study is first defined theoretically as a construct and then expressed in terms of observable behaviour (operationalization). Operationalization makes empirical research possible. At the same time when results are interpreted research findings are linked back to constructs. Moving from an operationalization to a construct is always a bit of a leap. Construct validity of an experiment is high if this leap is justified and if the operationalization provides sufficient coverage of the construct. For example, in some research studies anxiety was measured by a fidgetometer, a specially constructed chair that registers movements at various points and so calculates the amount of "fidgeting". Subjects would be invited to the laboratory and asked to wait in a chair, not suspecting that the experiment has already started. The rationale is that the more anxious you are, the more you fidget in the chair. Are the readings of a fidgetometer a good operationalization of anxiety? On the one hand, it is an objective measure. On the other hand, fidgeting may be a symptom of something other than anxiety. Also the relationship between anxiety and increased fidgeting first has to be demonstrated in empirical research.

Internal validity characterizes the methodological quality of the experiment. Internal validity is high when confounding variables have been controlled and we are quite certain that it was the change in the IV (not something else) that caused the change in the DV. In other words, internal validity links directly to bias: the less bias, the higher the internal validity of the experiment. Biases in the experiment (threats to internal validity) will be discussed below.

External validity characterizes generalizability of findings in the experiment. There are two types of external validity: population validity and ecological validity. **Population validity** refers to the extent to which findings can be generalized from the sample to the target population. Population validity is high when the sample is representative of the target population and an appropriate sampling technique is used. **Ecological validity** refers to the extent to which findings can be generalized from the experiment to other settings or situations. It links to the artificiality of experimental conditions. In highly controlled laboratory experiments subjects often find themselves in situations that do not resemble their daily life. For example, in memory experiments they are often asked to memorize long lists of trigrams. To what extent can findings from such studies be applied to everyday learning situations?

There is an inverse relationship between internal validity and ecological validity. To avoid bias and control for confounding variables, you make the experimental procedures more standardized and artificial. This reduces ecological validity. Conversely, in an attempt to increase ecological validity you may allow more freedom in how people behave and what settings they choose, but this would mean that you are losing control over some potentially confounding variables.


Validity
Internal
External
Construct
To what extent is the change in DV caused by IV?
Population
Ecological
To what extent do the operationalizations reflect the construct?
Credibility
To what extent can the findings be generalized to the wider population?
To what extent can the findings be generalized to real-life settings?
Generalizability (to theory)
Generalizability (to other people)
Generalizability (to other situations)

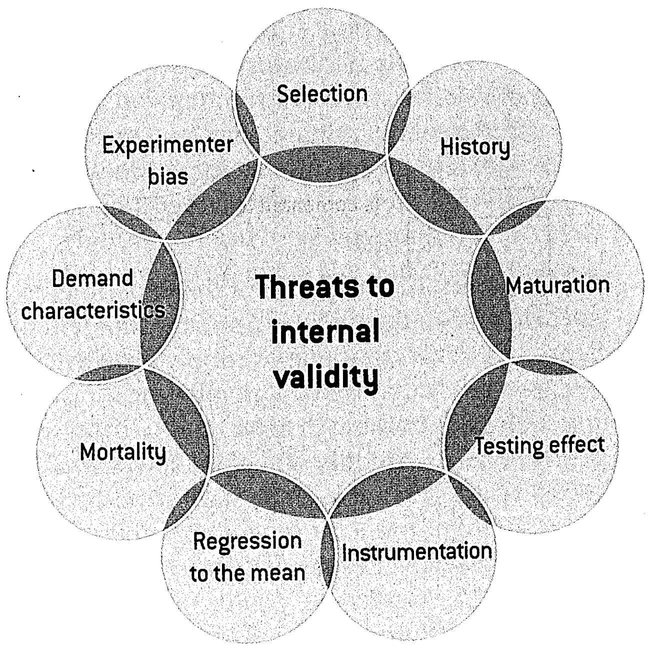
▲ Figure 1.5 Validity of experiments

Exercise

- Leaf through this book (consider the units on the biological, cognitive or sociocultural approach to behaviour), find a description of any experimental study and analyse its construct, internal and external validity. If you feel that you do not have enough detail, you could find more information on the study online, or even read the original article.
- Present the results of your analysis in class.

Bias in experimental research: threats to internal validity

Bias in experimental research comes in the form of confounding factors that may influence the cause-and-effect relationship between the IV and the DV, decreasing internal validity. Below you will find a description of several common sources of threat to internal validity, based on Campbell (1969).



▲ Figure 1.6 Sources of threat to internal validity

1. **Selection**. This occurs if for some reason groups are not equivalent at the start of the experiment: apart from the planned IV-related difference, they differ in some other variable. As a result, we cannot be sure if the post-experiment differences between groups reflect the influence of the IV or this other variable. Selection occurs in independent measures and matched pairs designs in case group allocation was not completely random.

2. **History**. This refers to outside events that happen to participants in the course of the experiment. These outside events become a problem when they can potentially influence the DV or are not evenly distributed in the comparison groups. History is especially important in lengthy experiments where the DV is measured sometime after the onset of the study. For an example of history-related bias think of a memory experiment where participants are required to memorize long lists of words and the experiment is conducted in two groups (experimental and control) simultaneously in two different rooms on the opposite sides of a school. As the experiment begins, there is some noise coming from road construction outside. The control group is closer to the construction site so the noise in their room is louder. Since distracting noise can affect memory performance and levels of noise were not equal in the two groups, resulting differences in the DV may reflect the influence of the IV as well as the confounding variable (noise). To counteract history as a threat to internal validity such confounding variables should be either eliminated or kept constant in all comparison groups (for example, change the rooms so that they are both on the same side of the school building).

3. **Maturation**. In the course of the experiment participants go through natural developmental processes, such as fatigue or simply growth. For example, suppose you are piloting a psychological training programme to increase assertiveness in middle school students. You measure assertiveness at the start, conduct the training programme for several months and measure assertiveness again. The resulting increase of assertiveness may be due to either the IV (the training) or simply to the fact that the middle school students grew up a little and naturally became more assertive. The counteracting strategy would be using a control group (the same time period, the same measurements but no training sessions).

4. **Testing effect**. The first measurement of the DV may affect the second (and subsequent) measurements. For example, suppose you are investigating the effectiveness of a video to reduce test anxiety in primary school children. For this your participants take an ability test preceded by a self-report anxiety measure at time A. They then watch your specially designed video and repeat the procedure (test and self-report anxiety measure) at time B. The difference in anxiety between time A and time B may be the result of both the video and the fact that it is their second time taking the test—they are more familiar with the format and therefore may be naturally less anxious. A solution to this is to use a control group where you show a neutral video of the same duration. Suppose you get the following results:

	Test anxiety (on a scale 0–100)	
Group	Before Test 1	Before Test 2
Experimental (specially designed video)	90	55
Control (neutral video)	90	70

▲ Table 1.5

Analysis of these results can reveal that a reduction of anxiety by 20 points is probably due to the testing effect; however, over and above that there is a 15-point anxiety effect of the specially designed video.

In repeated measures designs testing effect is a special case of order effects, and counterbalancing is used to control for it.

5. **Instrumentation**. This effect occurs when the instrument measuring the DV changes slightly between measurements. For psychology this becomes relevant when you consider that an "instrument of measurement" is often a human observer. Suppose you are investigating bullying on a school campus during breaks. You are looking at two groups of students who are exposed to different experimental conditions. If you observe group 1 in the morning and group 2 in the afternoon, you might be more tired in the afternoon and miss some important behaviours. If you observe one of the groups during a short break and the other one during the lunch break, observations during the lunch break may be less accurate because it is more crowded. To avoid this researchers should try to standardize measurement conditions as much as possible across all comparison groups and all observers.

6. **Regression to the mean**. This is an interesting source of bias that becomes a concern when the initial score on the DV is extreme (either low or high). Extreme scores have a purely statistical tendency to become more average on subsequent trials. Suppose you have designed anxiety reduction training for students. To test its effectiveness, you administer an anxiety questionnaire in a group of students and select a sample of students who have the largest score (for example, 80–100 on a 100-point scale). With these students you then conduct your training session and measure their anxiety again. Even if we assume that testing effects are not an issue, we would expect extremely anxious students to naturally become less anxious even without the training session. To put it more precisely, the probability that extremely anxious students will become even more anxious is less than the probability that they will become less anxious. This means that statistically a reduction of anxiety should be expected. A counter-measure is a control group with the same starting average anxiety level and measurements at the same point of time, but without the intervention.

7. **Experimental mortality**. This refers to the fact that some participants drop out during an experiment, which may become a problem if dropouts are not random. Suppose you are investigating the influence of emotion on ethical decision-making. For this you give your participants a number of scenarios of the type "Would you kill 1 person to save 1000?" In the control group the description of this "one person" is neutral, but in the experimental group this is someone they know personally, so there is more emotional involvement. You hypothesize that people will be less likely to be utilitarian in their decision-making when they are personally involved (note that this research would create distress among participants and so raises ethical issues; it is quite possible such a study would not be approved by the ethics committee). Suppose that several participants in the experimental group refuse to continue participation and drop out, more so than in the control group. Ethical issues aside, this presents a methodological issue as well: even if the two groups were equivalent at the start of the experiment, they may be non-equivalent now. There appears a confounding variable (sensitivity) which is disproportionally represented in the two groups. There is no reliable way to counteract experimental mortality other than designing experimental conditions in such a way that participants would not feel the need to drop out.

8. **Demand characteristics**. This refers to a situation in which participants understand the purpose of the experiment and change their behaviour subconsciously to fit that interpretation. In other words, they behave in ways that they think the experimenter expects. This can happen for various reasons, for example, participants may feel that they will somehow be evaluated and so behave in a socially desirable way. To avoid demand characteristics, deception may be used to conceal the true purpose of the study (however, deception raises ethical issues—see below). You can consider using post-experimental questionnaires to find out to what extent demand characteristics may have influenced the results (this strategy does not prevent demand characteristics but just estimates their impact). Note that in repeated measures designs demand characteristics are a larger threat because participants take part in more than one condition and so have greater opportunities to figure out or guess the aim of the study.

9. **Experimenter bias.** This refers to situations in which the researcher unintentionally exerts an influence on the results of the study, for example, the Clever Hans case discussed above. Existence of this bias was first rigorously supported by Rosenthal and Fode (1963). In this experiment rats were studied for their maze-running performance. Rats were split into two groups at random, but the laboratory assistants (psychology students) were told that one of the groups was "maze-bright" and

the other one was "maze-dull" and that this difference in ability was genetic. Laboratory assistants had to follow a rigorous and standardized experimental procedure in which rats were tested on their performance in learning the maze task. This was supposed to be an identical study conducted with identical rats, but results showed that the rats labelled "maze-dull" performed significantly worse than the ones labelled "maze-bright". It was concluded that the result was an artifact: it was caused by experimenter bias rather than any genuine differences between the groups of rats. Post-experiment investigations revealed that experimenter bias was not intentional or conscious. The results were induced by subtle differences in the way laboratory assistants handled the rats. For example, without realizing it, assistants handled "maze-bright" rats for slightly longer and so stress was more reduced for these rats than for "maze-dull" rats. A counter-measure against experimenter bias is using so-called **double-blind designs** where information that could introduce bias is withheld both from the participants and from the people conducting the experiment. The study of Rosenthal and Fode would have been double-blind if the laboratory assistants had not been told which group of rats had which label.

Exercise

Once again leaf through this book and find a description of any experimental study.

- To what extent was this experimental study susceptible to one of the sources of threat to internal validity? What does it tell you about credibility of the study?
- If you do not have enough detail, find more information on the study online, or even read the original article.
- Present the results of your analysis in class.

ATL skills: Self-management

Athabasca University has a great learning resource on threats to internal validity. One tutorial consists of two parts, where part 1 is the theoretical background and definitions and part 2 is a practical exercise involving the analysis of 36 hypothetical experiments.

If you want to practise identifying potential sources of bias in experiments, you can access the tutorial here: https://psych.athabascau.ca/open/validity/index.php

Quasi-experiments versus true experiments

Quasi-experiments are different from "true" experiments in that the allocation into groups is not done randomly. Instead some pre-existing inter-group difference is used. "Quasi" is a prefix meaning "almost". The major limitation of a quasi-experimental design is that cause-and-effect inferences cannot be made. This is because we cannot be sure of the equivalence of comparison groups at the start of the study: pre-existing differences in one variable may be accompanied by a difference in unexpected confounding variables.

Suppose your hypothesis is that anxiety influences test performance. You have an opportunity sample of high school students. An intuitively obvious way to test this hypothesis would be to administer an anxiety questionnaire, divide the sample into two groups (anxious and non-anxious) based on the results, and then model a testing situation and compare test performance. The IV in this study is anxiety (it has two levels) and the DV is test performance. However, the researcher does not really manipulate the IV in this study. Pre-existing differences in anxiety are used, so we cannot be sure that anxiety is the only variable that differs in the two groups. For example, it is possible that high school students with high levels of anxiety also tend to have unstable attention, and it is actually attention that influences test performance. The bottom line is that we will be able to conclude that "anxiety is linked to test performance", but strictly speaking we will not be able to say "anxiety influences test performance".

To test the "influence" hypothesis a true experiment would be required, so we would have to manipulate the IV. How can you manipulate anxiety? One example is splitting participants randomly into two groups and telling one of the groups that they should expect results of their college applications later today. Anticipation of these results would probably increase anxiety in the experimental group. Then the test can be given. (Note that such an experiment would have ethical issues since it involves major deception and creates distress among participants.)

Other examples of pre-existing differences are age, gender, cultural background and occupation. Formation of experimental groups based on these variables implies a quasi-experiment. Sometimes a "true" experiment cannot be conducted because it is impossible to manipulate the IV (for example, how do you manipulate age or gender?) so quasi-experiments are justified.

In the way they are designed (superficially) quasi-experiments resemble "true" experiments, but in terms of the possible inferences (essentially) they are more like correlational studies.

Field experiments and natural experiments

Field experiments are conducted in a real-life setting. The researcher manipulates the IV, but since participants are in their natural setting many extraneous variables cannot be controlled. The strength of field experiments is higher ecological validity as compared to experiments in a laboratory. The limitation is less control over potentially confounding variables so there is lower internal validity. An example of a field experiment is **Piliavin, Rodin and Piliavin's** (1969) subway study in which the researchers pretended to collapse on a subway train and observed if other passengers would come to help. To manipulate the IV, some researchers were carrying a cane (the cane condition) while others were carrying a bottle (the drunk condition).

Natural experiments, just like field experiments, are conducted in participants' natural environment, but here the researcher has no control over the IV—the IV occurred naturally. Ecological validity in natural experiments is an advantage and internal validity is a disadvantage owing to there being less control over confounding variables. Another advantage of natural experiments is that they can be used when it is unethical to manipulate the IV, for example, comparing rates of development in orphans that were adopted and in those who stayed in the orphanage. Since researchers do not manipulate the IV, all natural experiments are quasi-experiments.

Type of experiment	Independent variable	Settings	Can we infer causation?
True laboratory experiment	Manipulated by the researcher	Laboratory	Yes
True field experiment	Manipulated by the researcher	Real-life	Yes (but there may be confounding variables)
Natural experiment	Manipulated by the nature	Real-life	No
Quasi-experiment	Not manipulated; pre-existing difference	Laboratory or real-life	No

▲ Table 1.6

Exercise

Go online and find examples of quasi-experiments, natural experiments and field experiments in psychology.

Quantitative research: correlational studies

Inquiry questions

- What does it mean for two variables to correlate with each other?
- What should be avoided when interpreting correlations?
- Can two correlating variables be unrelated in fact?
- Can correlations show curvilinear relationships?

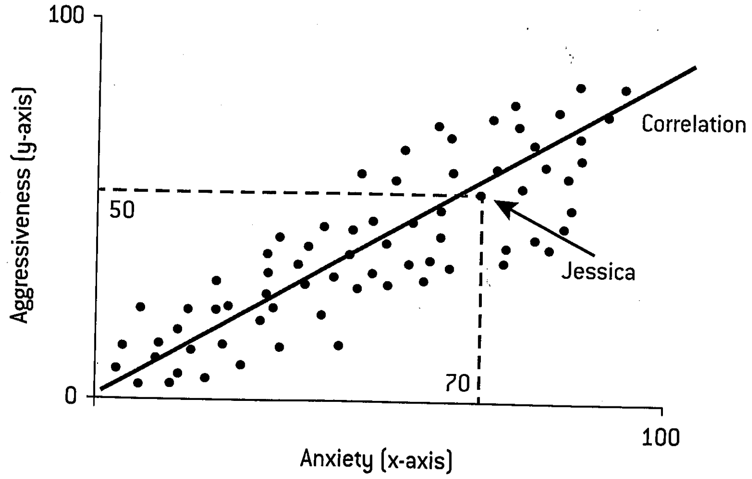
What you will learn in this section

- What is a correlation?
  - Effect size
  - Statistical significance
- Limitations of correlational studies
  - Causation cannot be inferred
  - The third variable problem
  - Curvilinear relationships
  - Spurious correlations
- Sampling and generalizability in correlational studies
- Credibility and bias in correlational studies

What is a correlation?

Correlational studies are different from experiments in that no variable is manipulated by the researcher, so causation cannot be inferred. Two or more variables are measured and the relationship between them is mathematically quantified.

The way it is done can be illustrated graphically through scatter plots. Suppose you are interested in investigating if there is a relationship between anxiety and aggressiveness in a group of students. For this you recruit a sample of students and measure anxiety with a self-report questionnaire and aggressiveness through observation during breaks. You get two scores for each participant: anxiety and aggressiveness. Suppose both scores can take values from 0 to 100. The whole sample can be graphically represented with a scatter plot.

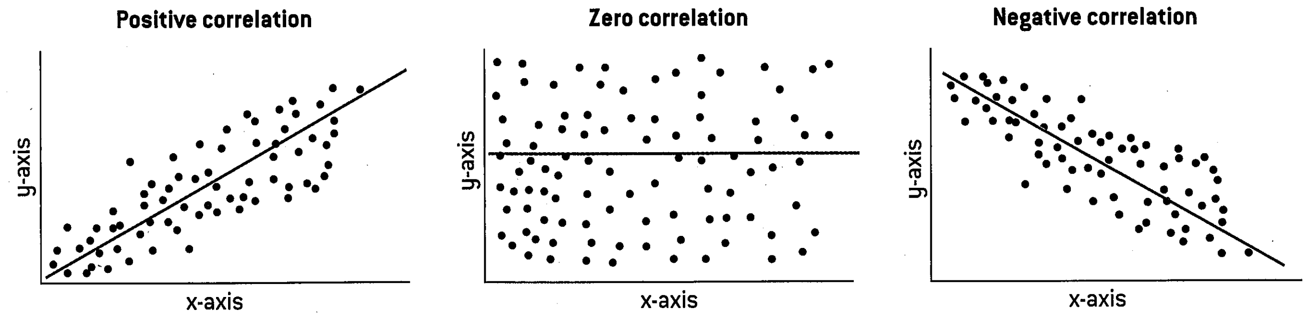


▲ Figure 1.7 Scatter plot

Each dot on the scatter plot represents one person. The coordinates of each dot give you the scores obtained for each of the variables. For example, Jessica's score on anxiety is 70 (the x-axis coordinate) and her score on aggressiveness is 50 (the y-axis coordinate). The whole scatter plot looks like a "cloud" of participants in the two-dimensional space of the two variables. A **correlation** is a measure of linear relationship between two variables. Graphically a correlation is a straight line that best approximates this "cloud" in the scatter plot.

In the example above, the correlation is positive because the cloud of participants is oblong and there is a tendency: as X increases, Y increases, so if an individual got a high score on variable X, that person probably also got a high score on variable Y, and vice versa. This is where the name "correlation" comes from: the two variables "co-relate". Remember that correlation does not imply causation: we cannot say that X influences Y, nor can we say that Y influences X. All we know is that there is a link between them.

A correlation coefficient can vary from −1 to +1. The scatter plots below demonstrate some examples:



▲ Figure 1.8 Examples of correlations

A positive correlation demonstrates the tendency for one variable to increase as the other variable increases. A negative correlation demonstrates the inverse tendency: when one variable increases the other variable decreases. The steeper the line, the stronger the relationship. A perfect correlation of 1 (or −1) is a straight line with the slope of 45 degrees: as one variable increases by one unit, the other variable increases (or decreases) by exactly one unit. A correlation close to zero is a flat line. It shows that there is no relationship between the two variables: the fact that a person scored high or low on variable X tells us nothing about his or her score on variable Y. Graphically such scatter plots are more like a circle or a rectangle.

Effect size and statistical significance

The absolute value of the correlation coefficient (the number from −1 to 1) is called the **effect size**. How do you know if a correlation is small or large? There are widely accepted guidelines based on **Cohen's** (1988) suggestions to interpret the effect size of correlations in social sciences.

Correlation coefficient effect size (r)	Interpretation
Less than 0.10	Negligible
0.10–0.29	Small
0.30–0.49	Medium
0.50 and larger	Large

▲ Table 1.7 Effect sizes for correlation coefficients

The effect size is not the only parameter that is important when interpreting a correlation coefficient. Another is the level of **statistical significance.** Statistical significance shows the likelihood that a correlation of this size has been obtained by chance. In other words, what is the probability that you will replicate the study with a different sample and the correlation will turn to zero? It depends on the sample size: with small samples you cannot be sure that an obtained correlation, even if it is relatively large, has not been obtained due to random chance. With large samples correlation estimates are more reliable and you can be more confident that the correlation is not a product of random chance but a genuine reflection of a relationship between the two variables in the population. The probability that a correlation has been obtained due to random chance can be estimated. Again, there are conventional cut-off points when results are considered to be "statistically significant" or not.

The probability that the result is due to random chance	Notation	Interpretation
More than 5%	p = n.s.	Result is non-significant
Less than 5%	p < .05	Result is statistically significant (reliably different from zero)
Less than 1%	p < .01	Result is very significant
Less than 0.1%	p < .001	Result is highly significant

▲ Table 1.8

The conventional cut-off point for statistical significance is 5%. Whatever result you obtained, if the probability that this result is pure chance occurrence is less than 5%, we assume that the result is statistically significant, reliably different from zero and so would be replicated in at least 95 out of 100 independent samples drawn from the same target population.

TOK

As you see, the nature of knowledge in psychology, just like the other social sciences, is probabilistic. We only know something with a degree of certainty and there is a possibility this knowledge is a product of chance.

How does that compare to the nature of knowledge in other areas such as natural sciences (physics, chemistry, biology), ethics or indigenous knowledge systems?

What can we do to increase the degree of certainty in social sciences (for example, think about replication of studies)?

When interpreting correlations one needs to take into account both the effect size and the level of statistical significance. If a correlation is statistically significant, it does not mean that it is large, because in large samples even small correlations can be significant (reliably different from zero). So, scientists are looking for statistically significant correlations with large effect sizes.

ATL skills: Research

Correlations are denoted by the letter r. Below are some examples of results of fictitious correlational studies. See if you can interpret them using your knowledge of Cohen's effect size guidelines and levels of statistical significance:

$r = 0.14, p = \text{n.s.}$

$r = 0.10, p < .05$

$r = 0.34, p < .01$

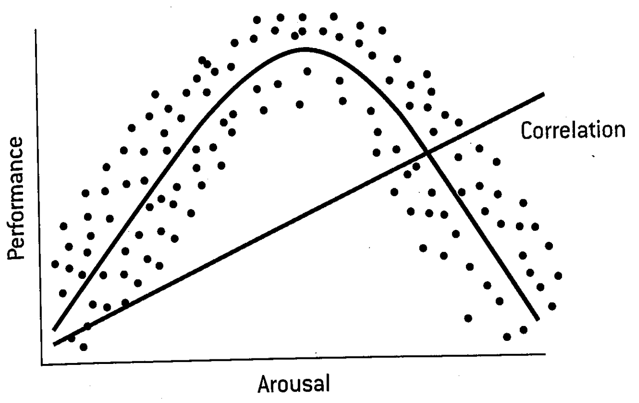
$r = 0.61, p < .001$

Limitations of correlational studies

Correlational studies have several major limitations.

- As already mentioned, correlations **cannot be interpreted in terms of causation**.
- **"The third variable problem"**. There is always a possibility that a third variable exists that correlates both with X and Y and explains the correlation between them. For example, cities with a larger number of spa salons also tend to have more criminals. Is there a correlation between the number of criminals and the number of spa salons? Yes, but once you take into account the third variable, the size of the city, this correlation becomes meaningless.
- **Curvilinear relationships**. Sometimes variables are linked non-linearly. For example, a famous Yerkes-Dodson law in industrial psychology states that there is a relationship between arousal and performance: performance increases as arousal increases, but only up to a point. When levels of arousal surpass that point, performance begins to decrease.

  Optimal performance is observed when levels of arousal are average. This can be seen in the scatter plot below.



▲ Figure 1.9 Arousal and performance

  However, this relationship can only be captured by looking at the graph. Since correlation coefficients are linear, the best they could do is to find a straight line that fits best to the scatter plot. So, if we were using correlational methods to find a relationship between arousal and performance, we would probably end up obtaining a small to medium correlation coefficient. Psychological reality is complex and there are a lot of potentially curvilinear relationships between variables, but correlational methods reduce these relationships to linear, easily quantifiable patterns.

- **Spurious correlations**. When a research study involves calculating multiple correlations between multiple variables, there is a possibility that some of the statistically significant correlations would be the result of random chance. Remember that a statistically significant correlation is the one that is different from zero with the probability of 95%. There is still a 5% chance that the correlation is an artifact and the relationship actually does not exist in reality. When we calculate 100 correlations and only pick the ones that turned out to be significant, this increases the chance that we have picked spurious correlations.

Sampling and generalizability in correlational studies

Sampling strategies in correlational research are the same as in experiments. First the target population is identified depending on the aims of the study and then a sample is drawn from the population using random, stratified, opportunity or self-selected sampling.

Generalizability of findings in correlational research is directly linked with sampling and depends on representativeness of the sample. Again, this is much like population validity in experiments.

Credibility and bias in correlational research

Bias in correlational research can occur on the level of variable measurement and on the level of interpretation of findings.

On the level of measurement of variables, various biases may occur and they are not specific to correlational research. For example, if observation is used to measure one of the variables, the researcher needs to be aware of all the biases inherent in observation. If questionnaires are used to measure variables, biases inherent in questionnaires become an issue. The list goes on.

On the level of interpretation of findings, the following considerations represent potential sources of bias.

- Curvilinear relationships between variables (see above). If this is suspected, researchers should generate and study scatter plots.
- "The third variable problem". Correlational research is more credible if the researcher considers potential "third variables" in advance and includes them in the research in order to explicitly study the links between X and Y and this third variable.
- Spurious correlations. To increase credibility, results of multiple comparisons should be interpreted with caution. Effect sizes need to be considered together with the level of statistical significance.

ATL skills: Self-management

Go back to the overview table (Table 1.2). Compare and contrast sampling, generalizability, credibility and bias in correlational research with those in experimental research.

- In what aspects are the approaches different?
- In what aspects are they the same?
- Are there any aspects where the ideas are similar but the terminology differs?