1 Research in psychology

Inquiry questions

- What is scientific psychology?
- How can we tell if a research study is credible?
- How can we study subjective phenomena objectively?
- How is correlation different from causation?
- How is quantitative research different from qualitative research?

What you will learn in this section

- What is psychology?
  - Psychology is the scientific study of behaviour and mental processes
  - Science and non-science
  - Behaviour and mental processes
  - A study of non-human animals
  - What IB psychology is not
- Research methodology: quantitative and qualitative methods
  - Qualitative versus quantitative comparison
  - Types of quantitative research: experimental, correlational, descriptive
  - Types of qualitative research
- Sampling, credibility, generalizability and bias in research: an overview

What is psychology?

"Psychology is the scientific study of behaviour and mental processes." This is the definition we are going to use throughout this book. Although it is quite a short definition, there are a lot of implications in it. Let's try and uncover them one by one.

Psychology is the scientific study... This part of the definition excludes such areas as pop psychology, that is, simple and appealing explanations that are not backed up by empirical evidence. What makes a theory or a study scientific, or where is the line between science and non-science? This is largely a TOK question and you will return to it throughout the book, but here are some major points.

- It should be supported by empirical evidence and be based on this evidence.
- It should be falsifiable, that is, it should be possible for the theory or study to be proven wrong.
- There should be a history of independent attempts to test the theory or replicate the study.

TOK

Science versus non-science demarcation is one of the key topics in TOK. The following concepts are important in the discussion of demarcation criteria:

- empirical evidence
- falsification/falsifiability
- replication.

While reading this unit, take a note of examples that illustrate these three concepts.

Think of other similar examples from such areas of knowledge as human sciences, natural sciences and mathematics.

Exercise

Look at the following research questions and pick one that you find interesting:

1. Do children who watch more violent TV shows become more violent?
2. Does extrasensory perception exist?
3. Are women attracted to men by the smell of their body?
4. Is abuse experienced differently in heterosexual and gay relationships?
5. Are breathing exercises effective for reducing test anxiety?
6. What emotions do people experience when watching horror movies in a cinema?
7. Are people in arranged marriages happier than people who married by choice?

If you were to conduct a research study to answer the question that you picked, how would you go about it? Think about details such as who your participants would be, what they would be required to do, how you would measure results and how you would ensure that the results are believable.

▲ Figure 1.1 Wilhelm von Osten and Clever Hans

In the early twentieth century, under the influence of Charles Darwin's theory of evolution, the public was very interested in animal intelligence: if humans evolved from animals, animals must be at least partially intelligent, so what exactly are they capable of? The case of Clever Hans sparked a lot of interest. Hans was a horse. Its owner Wilhelm von Osten, a mathematics teacher, claimed that he had taught Hans to solve arithmetic problems (addition, subtraction, multiplication, division, fractions), read, spell and understand some German. Questions could be asked verbally or in writing, and Hans would respond by tapping his hoof a certain number of times. Von Osten exhibited the horse frequently and gained a lot of public attention. A special committee was formed in Germany (called the Hans Commission). They ran a series of tests and concluded that the performance was not a fraud. So Hans's abilities were officially recognized as phenomenal!

However, another independent investigation carried out later by Oskar Pfungst, a psychologist, yielded different results. It demonstrated that Hans could not actually perform mental operations such as multiplication, but the horse was very responsive to clues that were provided by unsuspecting humans. To arrive at these conclusions, Pfungst successively tested a number of alternative hypotheses.

1. What if spectators give the horse hints or clues? He tested the horse and the questioner in the absence of spectators, but the horse continued to solve tasks correctly anyway.
2. What if von Osten himself gives the horse some clues? Another questioner was used during several trials, but the horse's performance did not worsen.
3. What if something in the questioner gives the correct answer away and the horse can feel that? Blinders were used to test this hypothesis. It turned out that when Hans was wearing blinders responses (the number of hoof taps) were incorrect most of the time. So, it was something in the questioner after all.
4. Did the questioners consciously let the horse know the correct answer, though? Additional trials were organized so that the questioner either knew or did not know the answer to the questions. It turned out that Clever Hans could only answer the questions correctly when the questioner knew the answer in advance.

This changed the focus of research from the horse to the questioner. When Pfungst carried out his observations, it was concluded that questioners who knew the answers had a tendency to become more tense as the hoof tapping approached the correct answer which would be reflected in their posture and facial expressions without them

realizing it. This was probably the clue that the horse was using. This makes sense evolutionarily, as detection of small postural changes is important as a survival skill for horses in the wild. Clever Hans certainly was clever, but the nature of his abilities was not mathematical (Goodwin, 2010)!

ATL skills: Thinking

How does Pfungst's investigation illustrate the concepts of empirical evidence, falsification and replication?

Von Osten himself, however, was never convinced of Pfungst's findings and he continued to exhibit the horse throughout Germany, gaining as much popularity as before. Nonetheless, scientifically, this was one of the starting points for designing rigorous experimental methodology in psychology and other human sciences. It was recognized that experiments, if not carefully controlled, could produce **artifacts**—results that are associated with the effect of unforeseen factors.

This whole story shows how claims can and should be tested scientifically, that is, by conducting a systematic evidence-based investigation that puts forward one hypothesis after another and tests them in a rigorous fashion. Note also how the whole investigation attempted to falsify the existing theory rather than support it.

... study of behaviour and mental processes. A scientific investigation requires an empirical approach to research, that is, relying on observation as a means of data collection. On the other hand, psychology (which comes from the Greek *psyche* = soul and *logos* = study, "the study of the soul") concerns itself with a wealth of phenomena, many of which are not directly observable. The first step in solving this dilemma is to identify something that can be observed directly. That's **behaviour**. Behaviour is everything that can be registered by an independent observer: it includes overt actions as well as gestures, facial expressions, verbal responses, endocrine reactions and so on. What stays "behind the scene" are the **mental processes** such as attention, perception, memory and thinking. We cannot observe them directly (which led some psychologists to say that they represent a "black box" and cannot be studied scientifically), but we can observe the indirect effects mental processes have on one's behaviour. So, we can infer something about the mental world as well.

ATL skills: Thinking

Brainstorm some behavioural indicators of the following:

- attention
- anxiety
- embarrassment.

To what extent do you think is it possible to use behavioural indicators to infer these "internal" phenomena? Would the inference be reliable?

Throughout this book we will use the term "behaviour" to refer to external, observable manifestations while the term "mental processes" will be used to denote internal patterns of information processing. However, you need to be aware of the fact that the term "behaviour" is often used in a more general sense, as an umbrella term for everything psychological. So sometimes you will encounter references to mental processes as types of "behaviour". This is not exactly accurate, but acceptable.

Note that the definition of psychology does not specify human behaviour or mental processes. This is because research with non-human animals is also an integral part of psychology. Since humans are just a stage in the continuous process of evolution, the study of animals may inform our understanding of human behaviour (and mental processes).

IB psychology is an academic discipline with an emphasis on rigorous research and scientific knowledge, but psychology is broader than pure academics and research. When people think about psychology many imagine counsellors and psychotherapists, practitioners who work with individual clients. University workers in lab coats conducting research is not the first thing that comes to mind. However, IB psychology focuses on academic knowledge and scientific research rather than counselling skills. This is because thorough understanding of psychological concepts and being able to think critically about psychological phenomena is of paramount importance in all spheres of psychology.

including counselling. It makes perfect sense to start with building these skills, much like the need to study aerodynamics before you are allowed to pilot an airplane.

Research methodology: quantitative and qualitative methods

All research methods used in psychology can be categorized as either quantitative or qualitative. Data in quantitative research comes in the form of numbers. The aim of quantitative research is usually to arrive at numerically expressed laws that characterize behaviour of large groups of individuals (that is, universal laws). This is much like the aim of the natural sciences in which it has been the ideal for a long time to have a set of simple rules that describe the behaviour of all material objects throughout the universe (think about laws of gravity in classic Newtonian physics, for example). In philosophy of science such orientation on deriving universal laws is called the **nomothetic approach**.

Quantitative research operates with **variables**. A variable ("something that can take on varying values") is any characteristic that is objectively registered and quantified. Since psychology deals with a lot of "internal" characteristics that are not directly observable, they need to be **operationalized** first. For this reason, there's an important distinction between **constructs** and **operationalizations**.

A construct is any theoretically defined variable, for example, violence, aggression, attraction, memory, attention, love, anxiety. To define a construct, you give it a definition which delineates it from other similar (and dissimilar) constructs. Such definitions are based on theories. As a rule constructs cannot be directly observed: they are called constructs for a reason—we have "constructed" them based on theory.

To enable research, constructs need to be operationalized. Operationalization of a construct means expressing it in terms of observable behaviour. For example, to operationalize verbal aggression you might look at "the number of insulting comments per hour" or "the number of swear words per 100 words in the most recent Facebook posts". To operationalize anxiety you might look at a self-report score on an anxiety questionnaire, the level of cortisol (the stress hormone) in the bloodstream or weight loss. As you can see, there are usually multiple ways in which a construct may be operationalized; the researcher needs to use creativity in designing a good operationalization that captures the essence of the construct and yet is directly observable and reliably measurable. As you will see throughout examples in this book, it is often a creative operationalization that makes research in psychology outstanding.

ATL skills: Research and communication

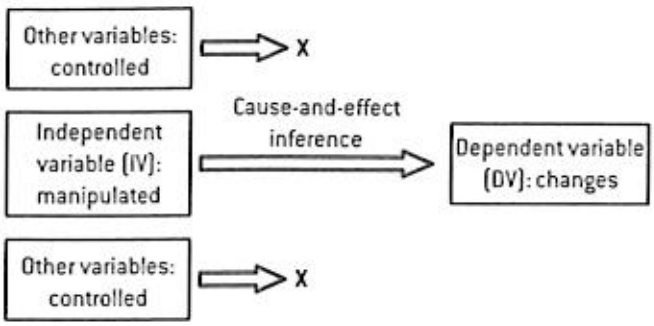
In small groups think of operationalizations of the following constructs: belief in God, assertiveness, shyness, pain, love, friendship, prejudice, tolerance to uncertainty, intelligence, wisdom.

Is it equally easy to operationalize them?

Discuss each other's operationalizations and outline their strengths and limitations.

There are three types of quantitative research.

- **Experimental studies**. The experiment in its simplest form includes one **independent variable** (IV) and one **dependent variable** (DV), while the other potentially important variables are controlled. The IV is the one manipulated by the researcher. The DV is expected to change as the IV changes. For example, if you want to investigate the effect of psychotherapy on depression, you might randomly assign participants to two groups: the experimental group will receive psychotherapy while the control group will not. After a while you might measure the level of depression by conducting a standardized clinical interview (diagnosis) with each of them. In this case the IV is psychotherapy. You manipulate the IV by changing its value: yes or no. The DV is depression; it is operationalized through the standardized diagnostic procedure. If the DV is different in the two groups, you may conclude that a change in the IV "caused" a change in the DV. This is why the experiment is the only method that allows **cause-and-effect inferences**.



▲ Figure 1.2 Cause-and-effect inference

- **Correlational studies.** Correlational studies are different from experiments in that the researcher does not manipulate any variables (there are no IVs or DVs). Variables are measured and the relationship between them is quantified. For example, if you want to establish if there is any relationship between violent behaviour of adolescents and how much time they spend watching violent television shows, you may recruit a sample of adolescents and measure their violent behaviour (by self-report, by ratings from classmates or even by observation in a natural setting) and the average number of hours per day spent watching violent television shows. Then you can correlate these two variables using a formula. Suppose you obtained a large positive correlation. This means that there's a trend in the data: the more time an adolescent spends watching violent shows, the more violent he or she is. However, you cannot make cause-and-effect inferences from correlational studies. Since you did not manipulate one of the variables, you do not know the direction of influence. It could be the case that watching violence influences violent behaviour (this would probably be the most popular, intuitive assumption). However, it is also possible that adolescents who behave violently choose to watch violent television programmes. Or there could even be a third variable (for example, low self-esteem) that influences both violent behaviour and watching violence on television. What you observe "on the surface" is just that—"co-relation", the fact that one variable changes as the other one changes.

ATL skills: Communication and social

In small groups come up with results of fictitious studies that would demonstrate either correlation or causation. Here are two examples.

1. In a group of adults we measured their attitudes to horror films and the number of siblings they have. We found that the more siblings you have, the more you like horror films.
2. We told one group of astronauts that their mission would start in a month and the other group that the mission would start in a year. We measured anxiety and found that it was higher in the group of astronauts who expected the mission to start in a month.

As you go through your list of fictitious studies, the other groups will have to say whether the study shows correlation or causation.

- **Descriptive studies.** In descriptive studies relationships between variables are not investigated, and the variables are approached separately. An example of a descriptive quantitative study would be a public opinion survey. We ask questions (for example, "Do you support the current policies of the President?") and we are interested in the distribution of answers to this particular question. Descriptive studies are often used in sociology and they are sometimes used in psychology to conduct a broad investigation of a phenomenon before "delving deeper" into the specifics.

Qualitative research is different. Its main focus is an in-depth study of a particular phenomenon. "In-depth" entails going beyond what can be objectively measured and quantified into the realm of human experiences, interpretations and meanings. Qualitative research makes use of such data collection methods as interviews or observations. Data comes in the form of texts: interview transcripts, observational notes, and so on. Interpretation of data involves a degree of subjectivity, but analysis is deeper than we can usually achieve through quantitative approaches. In philosophy of science such

orientation on an in-depth analysis of a particular case or phenomenon (without trying to derive universally applicable laws) is called the **idiographic approach**.

Parameter	Quantitative research	Qualitative research
Aim	Nomothetic approach: derive universally applicable laws	Idiographic approach: in-depth understanding of a particular case or phenomenon
Data	Numbers	Texts
Focus	Behavioural manifestations (operationalizations)	Human experiences, interpretations, meanings
Objectivity	More objective (the researcher is eliminated from the studied reality)	More subjective (the researcher is included in the studied reality)

▲ Table 1.1 Quantitative versus qualitative research

Qualitative research methods that we will discuss in this chapter are:

- observation
- interview
- focus group
- case study
- content analysis.

Sampling, credibility, generalizability and bias in research

Sampling, credibility, generalizability and bias are some of the characteristics used to describe a research study and make a judgment of its quality. These characteristics are universal for social sciences, but they can be approached very differently by quantitative and qualitative researchers, sometimes even with distinctly different sets of terms to express the same ideas. So it is important that you understand both these overarching concepts and the way they are broken down in quantitative as compared to qualitative research. Let's start with the overarching concepts.

A sample is the group of individuals taking part in the research study. **Sampling** is the process of finding and recruiting individuals for the study. There are different sampling techniques, and it is important to be aware of their strengths and limitations as sampling may affect the results of the study. For example, if the aim of your research is to see if anxiety correlates with aggression in teenagers (in general), but you only sample teenagers from one school in a criminal neighbourhood, your sampling technique will have important implications for the conclusions you will be able to make. Similarly, if you study political views of unemployed people and you recruit your sample by asking a small number of participants to bring their friends (and possibly friends of friends), you might end up with a limited sample because people of similar political views are more likely to be friends with each other.

Credibility refers to the degree to which the results of the study can be trusted to reflect the reality. It is closely linked to **bias**, because the results of the study do not reflect reality if there was some sort of bias in it. There are a lot of "traps" that a researcher can walk into. For example, in an interview, while the researcher believes the interviewee's responses to be true, participants may actually guess the aim of the study and respond in a way that they think the researcher is expecting them to. Or researchers themselves, being interested in confirming their hypothesis, may selectively notice supporting evidence and unintentionally ignore contradicting evidence. If there is indication that potential sources of bias were, to the best of our knowledge and abilities, controlled or eliminated, credibility of the research study is believed to be high. Quantitative and qualitative research approaches to credibility and bias are distinctly different, although they overlap in some aspects.

Generalizability refers to the extent to which the results of the study can be applied beyond the sample and the settings used in the study itself. Sometimes, especially in quantitative research, you want to generalize findings from the sample to a much wider group of people (called "population") because your aim is to discover universal laws of behaviour. Sometimes the research study is conducted in artificial settings

(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

could be reasons to believe that people are not that different in terms of the phenomenon under study. For example, if you study the influence of caffeine on attention, there are reasons to believe that results will be similar cross-culturally, and it might be a waste of time to use a stratified or a random sample. Finally, convenience sampling is useful when wide generalization of findings is not the primary goal of your research, for example, if you are conducting an exploratory study and you are not sure the hypothesis will be supported by evidence. If the hypothesis will not "work" in a small sample, why waste time testing it in a representative sample? Or you are replicating someone else's research and your aim is to see if the universal law (that was discovered by this someone) will hold true in your specific sample, thus trying to falsify prior theory. The limitation of convenience sampling is, of course, lack of representativeness.

- **Self-selected sampling**. This refers to recruiting volunteers. An example of this approach is advertising the experiment in a newspaper and using the participants who respond to the advert. The strength of self-selected sampling is that it is a quick and relatively easy way to recruit individuals while at the same time having wide coverage (many different people read newspapers). The most essential limitation, again, is representativeness. People who volunteer to take part in experiments may be more motivated than the general population, or they may be looking for the incentives (in many studies participants are financially rewarded for their time).

Exercise

Now that you know what sampling strategies can be used in an experiment, how would you change your approach to recruiting a sample for the investigation of the influence of praise on school performance of teenagers?

Experimental designs

Experiments always involve manipulating some variables and measuring the change in others. But the specific ways in which this can be organized differ depending on the aims of the research. The organization of groups and conditions in an experiment is known as the experimental design, and there are three basic types of experimental design.

Independent measures design involves random allocation of participants into groups and a comparison between these groups. In its simplest form, you randomly allocate participants from your sample into the experimental group and the control group. Then you manipulate the experimental conditions so that they are the same in the two groups except for the independent variable. After the manipulation you compare the dependent variable in the two groups.

ATL skills: Research

Consider the difference between random sampling (selecting the sample from the target population) and random group allocation (dividing your sample into groups). It is possible to have random group allocation in non-random samples and vice versa.

The rationale behind random group allocation is that all potential confounding variables cancel each other out. If the groups are not equivalent at the start of the experiment, you will be comparing apples to oranges. Imagine that you are testing the hypothesis that praise at school improves students' performance and for this you take two existing groups of students, with one being rarely praised by their teachers and the other one often praised. Arguably, the groups might not be equivalent: they have different experiences with the teachers, different ingroup values and habits, and so on—but to account for all these potentially important factors is impossible.

Conversely, when the group sizes are sufficiently large and allocation is completely random, chances are that groups will be equivalent—the larger the sample, the higher the chance.

Of course, there could be more than two groups, depending on how many IVs you use and how many levels each variable has. In the above example, you could use more than one IV: the influence of praise and the allocation of homework on school performance. With two levels for each of these IVs you would need to randomly allocate participants into four groups:

	Homework given	Homework not given
Rarely praised	1	2
Frequently praised	3	4

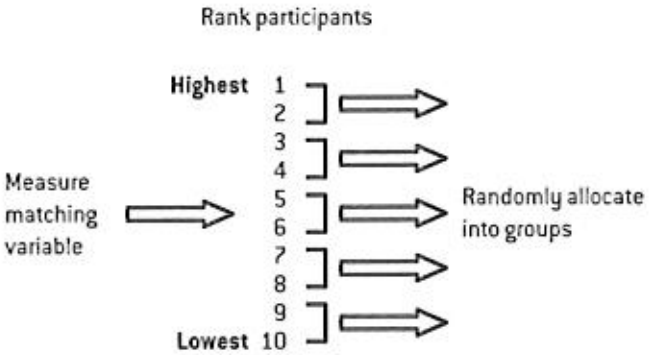
▲ Table 1.4

This experimental design with two IVs, each with two levels, is quite frequently used in psychological experiments. It is known as a 2 × 2 experimental design. Of course you can think of other combinations: 2 × 3 (two IVs, three levels in each), 3 × 2 (three IVs, two levels in each), 4 × 4 (four IVs, four levels in each). The more cells you have in this table, the larger the sample you need, so at some point it becomes impractical to increase the number of groups.

To summarize, regardless of the number of IVs and their levels, an experiment follows an independent measures design when the IV is manipulated by randomly allocating participants into groups. This allows us to assume that the groups are equivalent from the start so whatever difference we observe at the end of the experiment must have been caused by our experimental manipulation.

Matched pairs design is similar to independent measures. The only difference is that instead of completely random allocation, researchers use matching to form the groups.

To illustrate matching, let's consider an example. Suppose you are conducting a study of the effect of sleep deprivation on memory. For this you need two groups of participants. One of the groups will sleep peacefully in the laboratory and the other group will be woken up every 15 minutes. In the morning you will give both groups a memory test and compare their performance. You suspect that there is one confounding variable that may influence the results: memory abilities. Some people generally have better memory than others, therefore it is important to you that the two groups at the start of the experiment are equivalent in their memory abilities. Random allocation will usually make that happen, but you only have 20 participants (10 in each group). With a small sample like this there is a chance that random allocation will not work. So you want to control the equivalence of memory abilities "manually" while leaving everything else to random chance. For this you test memory abilities in your participants prior to the experiment. Then you rank participants according to their memory abilities (for example, from the highest to the lowest). Then you take the first two participants from the top of the list and randomly allocate one of them to the experimental group and the other one to the control group. You take the next two participants and repeat the procedure for the rest of the list. The two resulting groups are certainly equivalent in terms of memory abilities and probably (due to random chance) equivalent in all other characteristics.



▲ Figure 1.3 Matched pairs design

The variable that is controlled (memory abilities in the example above) is called the **matching variable**. Matched pairs designs are preferred when:

- the researcher finds it particularly important that the groups are equivalent in a specific variable
- the sample size is not large, therefore there is a chance that random allocation into groups will not be sufficient to ensure group equivalence.

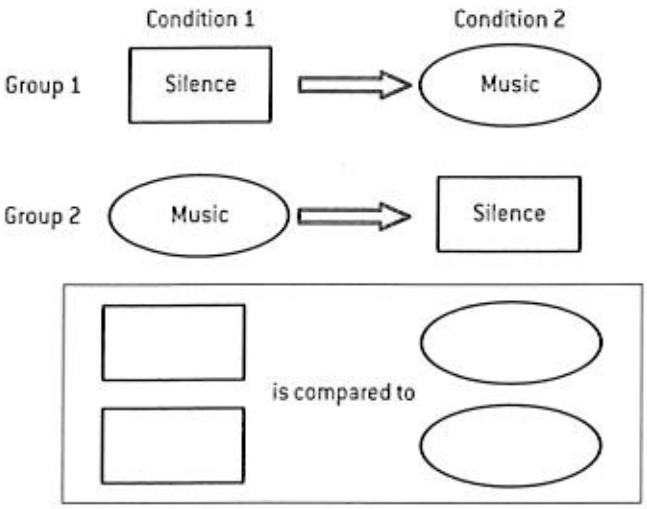
Repeated measures design is used when the goal is to compare conditions rather than groups of participants. The same group of participants is exposed to two (or more) conditions, and the conditions are compared. For example, imagine your aim is to investigate the effect of classical music on learning. You ask your participants to learn a list of trigrams (meaningless combinations of three letters such as HPX, LJW) for 10 minutes in silence and register the number of trigrams correctly recalled. Then you ask the same participants to learn a different list of trigrams for another 10 minutes, but this time with classical music playing in the background. You compare results from the first and the second trial.

The problem with repeated measures designs is that they are vulnerable to **order effects**: results

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.

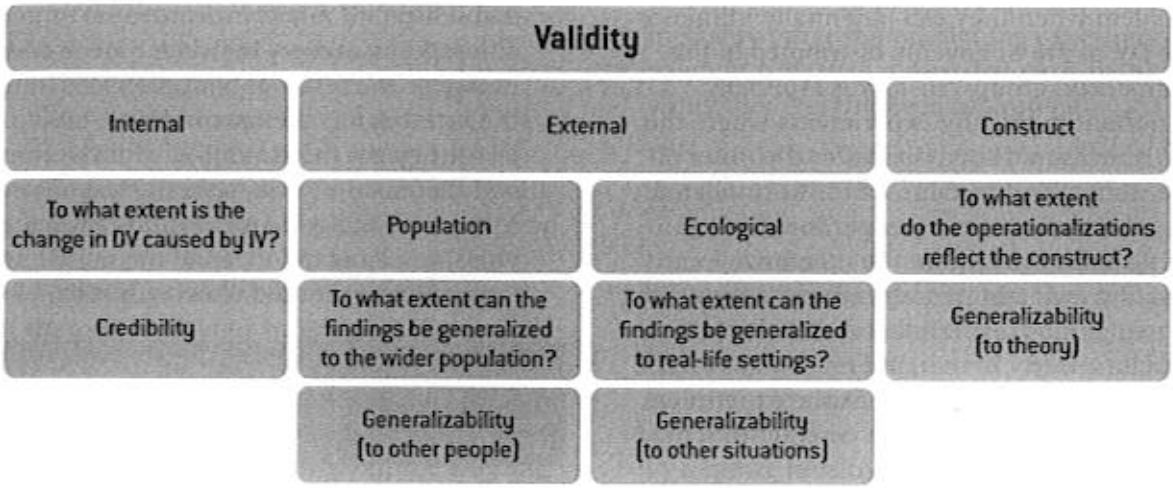
Construct validity characterizes the quality of operationalizations. As you know, the phenomenon under study is first defined theoretically as a construct and then expressed in terms of observable behaviour (operationalization). Operationalization makes empirical research possible. At the same time when 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.



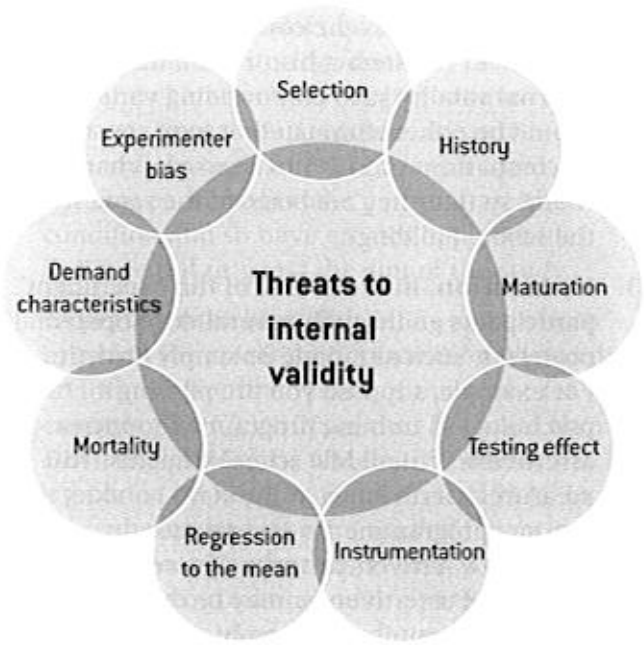
▲ Figure 1.5 Validity of experiments

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?

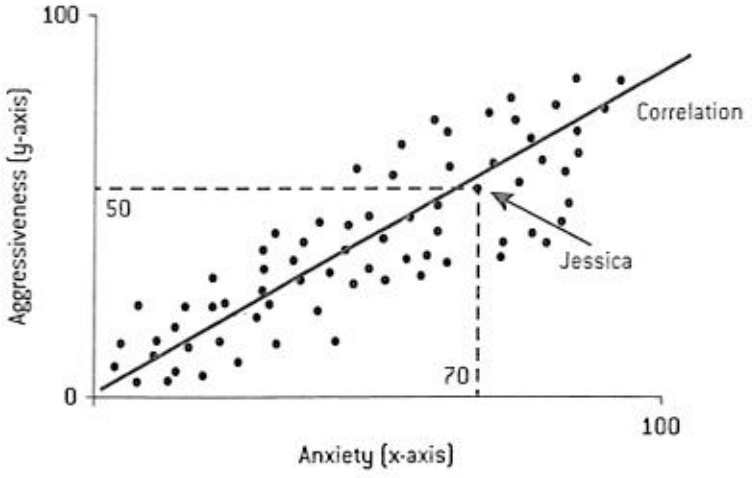
What you will learn in this section

- 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.

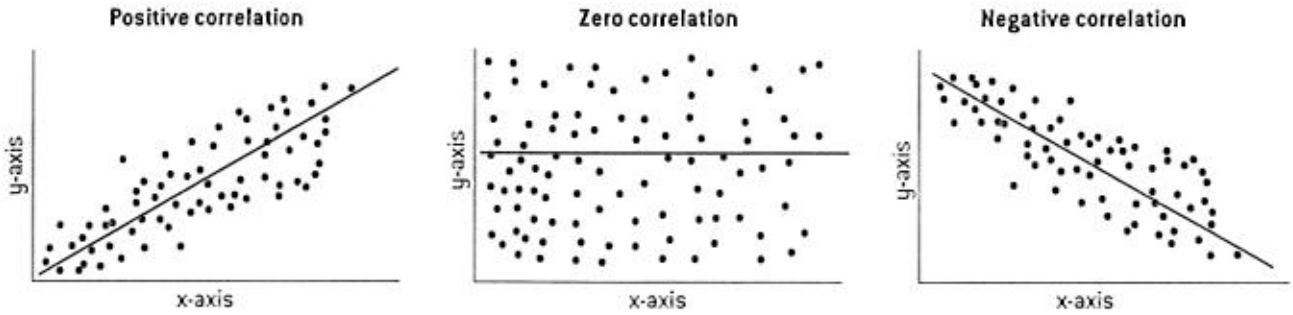


▲ Figure 1.7 Scatter plot

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:

$r = 0.14, p = n.s.$

$r = 0.10, p < .05$

$r = 0.34, p < .01$

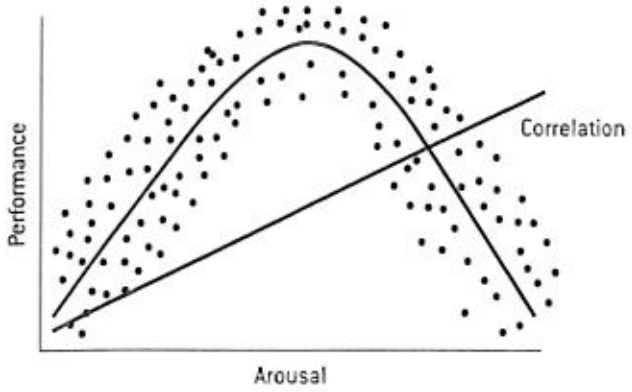
$r = 0.61, p < .001$

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?

Qualitative research

Inquiry questions

- To what extent can findings from qualitative research be generalized?
- How can credibility of qualitative research studies be ensured?
- What are the differences and similarities in how qualitative and quantitative research approaches sampling, credibility, generalizability and bias?

What you will learn in this section

- Credibility in qualitative research
  - Triangulation: method, data, researcher, theory
  - Rapport
  - Iterative questioning
  - Reflexivity: personal, epistemological
  - Credibility checks
  - Thick descriptions
- Bias in qualitative research
- Participant bias
  - Acquiescence bias
  - Social desirability bias
  - Dominant respondent bias
  - Sensitivity bias
- Researcher bias
  - Confirmation bias
  - Leading questions bias
  - Question order bias
  - Sampling bias
  - Biased reporting
- Sampling in qualitative research
  - Quota sampling
  - Purposive sampling
  - Theoretical sampling
  - Snowball sampling
  - Convenience sampling
- Generalizability in qualitative research
  - Sample-to-population generalization
  - Theoretical generalization
  - Case-to-case generalization = transferability

Credibility in qualitative research

Credibility in qualitative research is an equivalent of internal validity in the experimental method. As you have seen, internal validity is a measure of the extent to which the experiment tests what it is intended to test. To ensure internal validity in experimental research we need to make sure that it is the IV, not anything else, that causes the change in the DV. To do this, we identify all the possible confounding variables and control them, either by eliminating them or by keeping them constant in all groups of participants.

In a similar fashion, credibility in qualitative research is related to the question, "To what extent do the findings reflect the reality?" If a true picture of the phenomenon under study is being presented, the study is credible.

The term "**trustworthiness**" is also used to denote credibility in qualitative research.

TOK

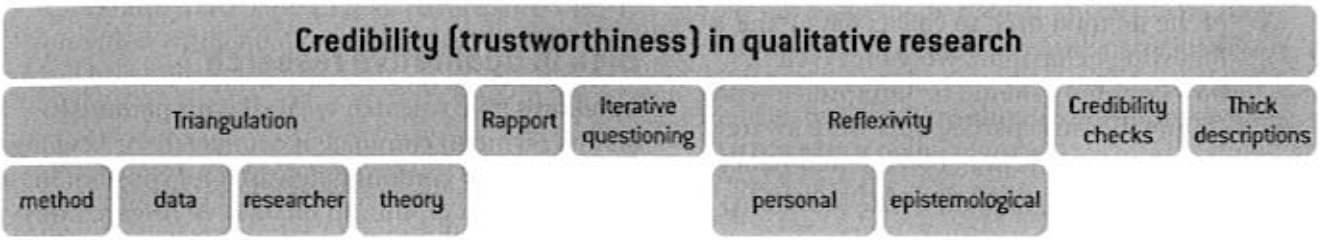
How do we know if the picture of a phenomenon presented in the findings from a qualitative study is "true"? If we had a way to know that, we wouldn't need a research study in the first place!

One of the popular definitions of knowledge is "justified true belief". A similar problem, however, arises with this definition: other than through "knowledge", we do not have a way of establishing if something is true. So, knowledge depends on truth but truth is a result of knowledge.

To solve this paradox, it has been suggested to substitute "true" in this definition to "beyond reasonable doubt". So, to ensure that a qualitative research study is credible we need to demonstrate that its findings are "true beyond reasonable doubt".

How do you understand that? What do you think is "reasonable doubt" in this context?

To ensure that what is presented in the findings of a qualitative study is true, several types of measures can be taken.



▲ Figure 1.10 Trustworthiness

- **Triangulation**. This refers to a combination of different approaches to collecting and interpreting data. There are several types of triangulation all of which can be used to enhance the credibility of a study.
  - Method triangulation. The use of different methods in combination can compensate for their individual limitations and reinforce their strengths. If the same results are obtained using various methods (for example, interviews and observations), credibility increases.
  - Data triangulation. This refers to using data from a variety of accessible sources. For example, if participants during an interview refer to certain documents, these documents may be studied in order to gain a clearer understanding of the participants' experiences. Observations may be supported by studying documented biographical data, and so on.
  - Researcher triangulation. As follows from the name, this refers to combining observations/interpretations of different researchers. Undoubtedly, if two people see the same thing, this increases credibility of their findings.
  - Theory triangulation. This refers to using multiple perspectives or theories to interpret the data.
- **Establishing a rapport**. Researchers should ensure that participants are being honest. For example, the researcher should remind participants about voluntary participation and the right to withdraw so that responses are only obtained from participants who are willing to contribute. It should be made clear to participants that there are no right or wrong answers and in general a good rapport should be established with participants so that they alter their behaviour in the presence of the researcher as little as possible.
- **Iterative questioning**. In many research projects, especially those involving sensitive data, there is a risk that participants will distort data either intentionally (lying) or unintentionally to try to create a certain impression on the researcher. Spotting ambiguous answers and returning to the topic later while at the same time rephrasing

the question might help researchers to gain a deeper insight into the sensitive phenomenon.

- **Reflexivity**. Researchers should reflect on the possibility that their own biases might have interfered with the observations or interpretations. Arguably, due to the nature of qualitative research that requires the involvement of the researcher in the studied reality, a certain degree of bias is unavoidable. However, researchers need to be able to identify the findings that might have been affected by these biases the most, and if they were affected, how. There are two types of reflexivity:
  - **epistemological reflexivity**, linked to knowledge of the strengths and limitations of the method used to collect data ("the following behaviours were observed ... however, they should be interpreted with caution because participants were aware that they were being observed and hence might have modified their behaviour")
  - **personal reflexivity**, linked to the personal beliefs and expectations of the researcher ("I noticed that overcoming trauma was particularly emphasized in their conversations, however, since I myself have a history of overcoming childhood trauma, this observation could have been influenced by my personal beliefs and should be cross-checked by an independent interviewer").
- **Credibility checks**. This refers to checking accuracy of data by asking participants themselves to read transcripts of interviews or field notes of observations and confirm that the transcripts or notes are an accurate representation of what they said (meant) or did. This is often used in interviews with the interviewees receiving the transcripts or notes and being asked to correct any inaccuracies or provide clarifications.
- **"Thick descriptions"**. This refers to explaining not just the observed behaviour itself, but also the context in which it occurred so that the description becomes meaningful to an outsider who never observed the phenomenon first-hand. Essentially it boils down to describing the phenomenon in sufficient detail so that it can be understood holistically and in context. For example, imagine a stranger smiled at you. This behaviour out of context can be reported "thinly", just stating the fact, or it can be placed in a context (who, where, in what circumstances), making it meaningful. To provide thick descriptions researchers should reflect anything that they observe and hear including their own interpretations, even if some of these details do not seem significant at the time. Thick descriptions are also referred to as "rich" descriptions; these terms are interchangeable.

ATL skills: Research

To what extent is this similar to the way internal validity is ensured in experimental research? What are the differences?

Bias in qualitative research

In quantitative research we deal with potential bias by trying to eliminate it completely or keeping the potentially confounding variables constant in all comparison groups. In qualitative research this approach is not possible, and bias is actually an integral part of the research process because the researcher is a tool through which data is collected. So, while some types of bias may be avoided, other types of bias are inevitable and need to be reflected on and accounted for.

Sources of bias in qualitative research may be associated both with the researcher and the participant. Let's look at the major sources of bias.

Participant bias

- **Acquiescence bias** is a tendency to give positive answers whatever the question. Some people are acquiescent by nature, and in some others acquiescence may be induced by the nature of the questions or the researcher's behaviour. To avoid this bias, researchers should be careful not to ask leading questions, making their questions open-ended, neutral and focused on the opinions of the participant.
- **Social desirability bias** is participants' tendency to respond or behave in a way that they think will make them liked or accepted. Participants may guess (or at least have a vague idea about) the aim of the study and try to look better than they really are. This may be done intentionally or unintentionally. Research into sensitive topics is especially vulnerable to social

desirability. To reduce this bias, questions should be phrased in a non-judgmental way that suggests that any answer is acceptable. Another trick that researchers use is to ask questions about a third person (for example, what do your friends think about ...?). This helps participants to disengage from the sensitive topics and provide more honest answers.

- **Dominant respondent bias** occurs in a group interview setting when one of the participants influences the behaviour and responses of the others. Dominant respondents may "hijack" talking time or intimidate others by demonstrating their assertiveness or superior knowledge of the subject. Researchers should be trained to keep dominant respondents in check and make sure that all participants are provided with equal opportunities to speak and are in a safe and comfortable environment to voice their opinions.
- **Sensitivity bias** is a tendency of participants to answer regular questions honestly, but distort their responses to questions on sensitive subjects. They may even give incorrect information to hide secrets. The solution to this problem is to build a good rapport with each participant and create trust between the participant and the researcher. To build trust, the researcher needs to behave professionally, make ethical guidelines regarding issues such as confidentiality absolutely clear to the participant and increase the sensitivity of the questions gradually while being responsive to the participant's concerns.

Researcher bias

- **Confirmation bias** occurs when the researcher has a prior belief and uses the research in an unintentional attempt to confirm that belief. Confirmation bias may influence the way questions are worded, the small nuances in the researcher's non-verbal behaviour, and selectivity of attention while observing behaviour or interpreting the data. Information that supports the prior belief is attended to, while information that contradicts it is disregarded. Reflexivity is the solution to confirmation bias. Confirmation bias is such a deeply grounded error in human information processing that it is largely unavoidable in qualitative research where data can only be collected "through" a human observer. So rather than avoiding it, researchers should be trained to recognize it and take it into account. If the possibility of bias is recognized, research can then be repeated with another observer to corroborate the findings (or not).
- **Leading questions bias** occurs when respondents in an interview are inclined to answer in a certain way because the wording of the question encourages them to do so. Even if an interview is carefully planned in advance, researchers often ask additional follow-up or clarification questions, and these may potentially cause distortions in the responses. Interviewers should be rigorously trained in asking open-ended, neutral questions that do not suggest a particular answer. Also they should avoid paraphrasing the participant's response to make sure they understood it correctly. Questions should be worded in the participant's own language.
- **Question order bias** occurs when responses to one question influence the participant's responses to the following questions. This bias stems from the human tendency to be consistent in our beliefs and actions. For example, if the first question on the interview asked if you liked sports and you hesitated but said yes, you would probably be inclined later to give more positive answers about your attitudes to gym membership. To minimize this bias, general questions should be asked before more specific ones, positive questions before negative ones, and behaviour questions before attitude questions.
- **Sampling bias** occurs when the sample is not adequate for the aims of the research. For example, the selection of people who are not "the best fit" in terms of the research purposes may be the result of convenience sampling. Also there are "professional participants" who look for opportunities to take part in research that provides financial incentives for participation. Although they can be accessed quickly and recruited easily, samples consisting entirely of "professional participants" should be used with caution.
- **Biased reporting** occurs when some findings of the study are not equally represented in the research report. For example, the researcher

might choose to only briefly mention pieces of evidence that do not "fit". Reflexivity, integrity and training of researchers are the means to counteract biased reporting.

Bias in qualitative research

Participant bias	Researcher bias
• Acquiescence	• Confirmation bias
• Social desirability	• Leading questions bias
• Dominant respondent	• Question order bias
• Sensitivity	• Sampling bias
	• Biased reporting

▲ Figure 1.11 Types of bias in qualitative research

To sum up, some types of bias in qualitative research may be eliminated while some others need to be recognized and taken into account. Reflexivity and triangulation are the two most important instruments that allow the researcher to reduce the influence of bias in qualitative research.

With regards to researcher bias, special attention needs to be paid to incorporating all data in the report and acknowledging the limitations of the research study, as well as asking independent researchers to review the results and procedure followed. With regards to participant bias, it is important to ask carefully crafted, indirect and open-ended questions and maintain neutrality.

The presence of biases is directly linked to both credibility and generalizability of research findings.

ATL skills: Thinking and self-management

The sources of bias in experimental and qualitative research appear in the table below. See if you can find any overlaps and discuss in class.

Experimental research	Qualitative research
Selection	Acquiescence bias
History	Social desirability bias
Maturation	Dominant respondent bias
Testing effect	Sensitivity bias
Instrumentation	Researcher bias
Regression to the mean	Confirmation bias
Experimental mortality	Leading questions bias
Experimenter bias	Question order bias
Demand characteristics	Sampling bias
	Biased reporting

Sampling and generalizability in qualitative research

Generalization is a broad inference from particular observations. It is "an inference about the unobserved based on the observed" (Polit and Beck, 2010, Elsevier).

Traditionally generalizability has been the focus of debate between supporters of quantitative and qualitative methods. The main argument against generalizability in qualitative research is that samples are not statistically representative of the target population. As you know, representativeness in quantitative research is a necessary requirement for findings to be applied beyond the sample to the target population it represents. A "weak" counter-argument to that is to say that qualitative methods do not aim to apply research findings to a wider population, in other words, the purpose of qualitative methods is the study of a particular sample but not the population it "represents". However, some scholars make a stronger argument and claim that generalizability is in fact achievable, to a certain extent, in qualitative research.

There are other arguments too, less popular, but no less valid. Some scientists doubt that generalizability is possible in principle, even in quantitative studies. They argue that every research study is embedded in a certain context (sample, setting, time, and so on), and generalization of findings would always include a degree of unsubstantiated speculation. Some other scholars argue that qualitative research is in fact more generalizable. They claim that rich data obtained in qualitative studies allows us to gain a deeper understanding of the phenomenon and so make more accurate inferences about its nature.

Sampling

In quantitative research, representativeness of the sample (and therefore the ability to generalize results to a wider population) is ensured through random sampling. In random sampling each member of the target population has an equal chance of being included in the sample. In other words, random sampling is probabilistic. However, sampling in qualitative research is **non-probabilistic**. These are the

most commonly used types of sampling in qualitative research.

- **Quota sampling.** In quota sampling it is decided prior to the start of research how many people to include in the sample and which characteristics they should have. This decision is driven by the research question—researchers look for people whose experiences would most likely provide an insight into the topic. Using various recruitment strategies, researchers then recruit participants until the quotas are met. Quota sampling is similar to stratified sampling in quantitative research in that both the important participant characteristics and the necessary sample proportions are pre-defined.
- **Purposive sampling.** This is similar to quota sampling in the sense that the main characteristics of participants are defined in advance and then researchers recruit participants who have these characteristics. However, the proportions and the sample size are not defined.
- **Theoretical sampling.** This is a special type of purposive sampling that stops when the point of data saturation is reached. **Data saturation** means that no new information is obtained from new participants added to the sample. Whether information is "new" or not is defined on the basis of a background theory: if no new evidence (or counterevidence) for the claims of the theory emerges, data saturation is reached. Generalization in this case is made from the data to the theory.
- **Snowball sampling.** In this approach a small number of participants are invited and asked to invite other people they know who also are of interest for the purposes of the research. This approach is mostly used in pilot research studies (when there are insufficient resources to carefully select participants) or in research with groups of people who are very difficult to reach (for example, drug users, youth gang members).
- **Convenience sampling.** The most superficial approach where you just use the sample that is easily available or accessible (for example, professors might conduct research with university students simply because it is time- and cost-efficient).

Generalizability of research findings in qualitative research may depend on the type of sampling used—studies using quota, purposive or theoretical sampling are more generalizable.

Types of generalizability

Firestone (1993) distinguished between three types of generalizability that provide a convenient framework for comparing quantitative and qualitative studies.

1. **Sample-to-population generalization.** The researcher starts by identifying the target population and then selects a sample that is representative of this population. The best approach to achieve this is to use random sampling. The concept that is used to describe sample-to-population generalizability in experiments is "population validity" (part of "external validity"). Due to the non-probabilistic nature of samples in qualitative research, this type of generalization is difficult.
2. **Theoretical generalization.** Generalization is made from particular observations to a broader theory. In quantitative research theoretical generalization takes the form of construct validity: it is the leap we make from directly observable operationalizations to the unobservable construct. In qualitative research theoretical generalization is achieved through rigorous analysis and interpretation of research findings: we can generalize to a wider theory if data saturation was achieved, thick descriptions provided, analysis was in-depth and free of biases, and so on. Theory plays a much greater role in qualitative research than in quantitative.
3. **Case-to-case generalization**, also known as **transferability.** Generalization is made to a different group of people or a different setting or context. In qualitative research transferability is the responsibility of both the researcher and the reader of the research report. The researcher's responsibility is to ensure that thick descriptions are provided so that the reader has sufficient information and details about the context of the study. The reader's responsibility is to decide whether or not the context described in

the report is similar to a new situation (Polit and Beck, 2010). A rough and pretty distant equivalent of transferability in quantitative research would probably be "ecological validity" (another part of "external validity").

ATL skills: Research and self-management

Compare the sampling techniques used in experiments and in qualitative research studies. Use any kind of visual representation to demonstrate the results of this comparison and present it in class.

How are the three types of generalizability approached in experiments and qualitative research studies? Which of these do you think are better achieved in qualitative research as compared to experimental research?

Go back to the overview table (Table 1.2) and see if it reflects your current knowledge of generalizability.

Qualitative research methods

Inquiry questions

- What is the range of qualitative methods used in psychology?
- How and why should one qualitative method be chosen over the others?

What you will learn in this section

- Observation
  - Reasons for choosing observation as the method
  - Reflexivity in observation
  - Types of observation: laboratory versus naturalistic; overt versus covert; structured versus unstructured; participant observation
- Interview
  - Reasons for choosing the interview
  - Interview transcripts and interview notes
  - Structured, semi-structured and unstructured interviews
- Focus group
  - Reasons for choosing the focus group
  - Limitations of the focus group method
- Content analysis
  - Five steps of inductive content analysis
  - Grounded theory
- Case study
  - Why are case studies referred to as a separate method?
  - Reasons for choosing the case study
  - Limitations of the case study method

Observation

There are several common reasons for choosing the method of observation.

- The focus of the research is on how people interact, interpret each other's behaviour and act upon these interpretations in a natural setting. For example, if you observe a group of primary school children in a typical enrichment class you may understand a lot about their everyday school life. Most other research methods are artificial in the sense that they place the participant in a specially constructed research context.
- The researcher believes that meaningful knowledge in the research area cannot be generated without observation, for example, because it cannot be articulated. For example, if you want to gain an insight into the behaviour of your classmates during a fire drill at your school, it will probably be more meaningful to observe an actual fire drill than to conduct an interview and analyse verbal responses.
- Observation allows the researcher to become immersed deeply into the studied phenomenon, sometimes even becoming part of it. Arguably, this is a strength because you gain almost first-hand experiences.

Observation is "experiential" and the researcher is strongly involved in the process of data generation. All generated data is the product of his or her selective attention and interpretations. This makes reflexivity especially important.

So, the main advantage of observation is the ability to generate diverse data about the behaviour of

participants in a naturally occurring setting. The major limitation would be susceptibility to biases, so reflexivity and other methods of ensuring credibility and generalizability of qualitative research need to be used extensively.

There are several types of observation, and the particular type chosen will have broad implications in terms of credibility, reflexivity, generalizability and ethics.

- **Laboratory versus naturalistic observation.** Naturalistic observation is carried out in naturally occurring settings, that is, a place that has not been arranged for the purposes of the study. Sometimes naturalistic observation would be the only choice (for example, in situations where it is unethical to arrange settings for the behaviour of interest to occur). If you wanted to study inter-group discrimination and violence, it would be unethical to encourage violence in a research setting. However, you may observe naturally occurring violence. A drawback is that it may be time-consuming because the behaviour of interest only occurs at certain times.
- Observation may be overt or covert. **Overt observation** occurs when participants are aware of the fact that they are being observed. Clearly the ethics of this approach are a strength as participants give their informed consent, but there are methodological limitations—biases related to participant expectations. When people know that they are being observed, they can intentionally or unintentionally change their behaviour. In contrast, in **covert observation** the researcher does not inform the members of the group about the reasons for his or her presence. An advantage of covert observation is gaining access to groups that would not normally agree to participate in research (for example, socially isolated or violent groups). Another strength is the avoidance of participant bias—subjects do not know they are being observed, so they behave naturally. The ethics here are a disadvantage. Participants do not give their consent to take part in the study. One way to avoid this issue is to debrief participants after the observation session and ask for their consent prior to using the data for research purposes.
- **Participant observation**. In this method the observer becomes part of the observed group. For example, many anthropologists spend time living among members of an indigenous society in order to study their culture "from the inside". For a great example of this, watch the BBC documentaries *Tribe* and *Amazon* with Bruce Parry. The advantage of participant observations is that they allow the researcher to gain first-hand experiences with the phenomenon of interest, gaining valuable insights. However, the drawbacks include the risk that the observer will lose objectivity as he or she becomes too involved with the studied group of individuals. This may happen because the researcher begins to identify himself or herself with the group. Of course, there is also the ethical issue: if participants do not realize that one of the members of their group is in fact an observer collecting information, this may be ethically questionable, especially in sensitive research topics.
- **Structured versus unstructured observation**. In structured observation information is recorded systematically and in a standardized way. For example, structured observation may be conducted with a checklist of behaviours of interest where the observer is required to note the occurrence of these specific behaviours in pre-defined time intervals. Rosen, Carrier and Cheever (2013) conducted structured observations of the use of technology among school students. Observers were equipped with a checklist of behaviours related to the use of technology (using a browser, using a telephone, and so on) and they had to fill out this checklist minute-by-minute. Unstructured observations do not have a pre-defined structure and observers simply register whatever behaviours they find noteworthy. Note that structured observation operates with numbers rather than text, which may be sufficient to say that structured observation is a quantitative research method. However, it is still idiographic rather than nomothetic (see Table 1.1).

Exercise

- Suppose your aim is to study ways in which destructive cults brainwash their new members, and observation is your method. What type of observation would you use and why?
- Describe how you would set up your research procedure both in terms of preparation and the actual observation process.

Interview

In-depth interviews are one of the most popular qualitative research methods for several reasons.

- This may be the only way to get an insight into the nature of subjective experiences and interpretations. Since attitudes, values, patterns of interpretation and other subjective phenomena are unobservable, the most straightforward way to study them is to rely on the participants' verbal reports.
- Interviews may be used to understand the meanings participants attach to certain events and their points of view. Again, this is not directly achievable by most other methods.
- In-depth individual interviews are useful when the topic is too sensitive for people to discuss in a group setting.

Interviews are a very personal form of research because there is direct contact between the interviewer and the interviewee. At the same time, interviews can, and often do, touch upon sensitive topics such as coping with a terminal illness, experiencing phobias, daily routines related to internet addiction and drug use.

Interviewing techniques are driven by the goal of learning as much as possible about the interviewee's opinions and experiences. The interviewer tries to build a rapport with the participant and then engage the person by asking neutral and carefully phrased questions, listening carefully to his or her responses and asking follow-up questions. The interviewer is the main research instrument. Tiny nuances in verbal and non-verbal behaviour of the interviewer may affect the interviewee's responses. For example, it is common in everyday conversations to ask leading questions, but interviewers must avoid doing it. This is why interviewers receive intensive training.

Interview data comes in the form of an audio or video recording which is subsequently converted to an **interview transcript**. Sometimes data also includes **interview notes**, accompanying observations about the participant and the interview context. Transcripts are later coded and analysed in line with the aims of the research.

There are three types of interview, depending on how fixed the list and the sequence of the questions is.

- **Structured interviews** include a fixed list of questions that need to be asked in a fixed order. It is most useful when the research project involves multiple interviewers and it is essential that they all conduct the sessions in a similar way. This allows many participants to be interviewed and some comparisons to be made (for example, comparing responses from male and female participants, across age groups, across cultures).
- **Semi-structured interviews** do not specify an order or a particular set of questions. They are somewhat like a checklist: the researcher knows that certain questions must be asked, but beyond that he or she can ask follow-up questions to get clarifications. If it better fits the natural flow of the conversation, the researcher can change the question order. Semi-structured interviews are better suited for smaller research projects, but they are also more effective in studying the unique experiences of each participant.
- **Unstructured interviews** are mostly participant-driven, and every next question is determined by the interviewee's answer to the previous one. Of course, the researcher still has to keep in mind the overall purpose of the research and stay focused on exploring a particular topic. However, two different interviewees may end up getting very different questions.

Exercise

Suppose you are interested in studying the reasons why teenagers join criminal groups. You used snowball sampling techniques to recruit 10 participants. Would you use a structured, semi-structured or unstructured interview? Why?

What do you think are the factors that need to be considered in conducting an interview with teenage gang members?

Focus group

The focus group is a special type of semi-structured interview that is conducted simultaneously with a group of 6–10 people. The key factor is that participants are encouraged to interact with each other and the interviewer serves as a facilitator. Participants discuss responses to every question and react to each other's statements. This provides additional data because they use their own language, agree and disagree with each other, enrich each other's perspectives and demonstrate a variety of opinions. The focus group facilitator can observe group dynamics and make use of it by directing group members' interaction so that they stay focused on the research topic.

The advantages of a focus group include the following.

- It is a quick way to get information from several participants at the same time.
- It creates a more natural and comfortable environment than a face-to-face interview, ensuring less participant bias.
- It is easier to respond to sensitive questions when you are in a group.
- Multiple perspectives are discussed so a more holistic understanding of the topic is achieved.

However, there are several "new" limitations that come as a cost for including group dynamics into the research process.

- If one of the participants is especially dominant, this may distort the responses of the other participants (for example, if they feel a need to conform), and it is the facilitator's responsibility to ensure that each participant contributes freely to the conversation.
- It is more difficult to preserve anonymity and confidentiality.
- Focus groups are especially demanding in terms of sampling and creating interview transcripts.

Content analysis

Interview recordings need to be transcribed and then analysed—but how do you analyse a text in a systematic and rigorous way while minimizing researcher bias? The widely used approach to analysing texts produced by participants is known as **inductive content analysis**, or thematic analysis. The goal of inductive content analysis is to derive a set of recurring themes. When extracting the themes the researcher has to maintain a balance between description and interpretation in the sense that the text needs to be interpreted, but these interpretations must be backed up by evidence from the text.

TOK

What is the difference between induction and deduction? If you do not remember, look it up.

Inductive content analysis follows a series of steps (Elo and Kungäs, 2008).

1. Writing the transcript. There are two types of transcript: verbatim or post-modern. Verbatim transcripts are word-for-word accounts of everything the participant said. Post-modern transcripts include notes about the intonation, gestures and other non-verbal elements in the participant's behaviour.
2. Reading the raw material several times and identifying initial themes. This is done iteratively. Researchers start with low-level themes, trying to stay as close to the text as possible. When the first reading is done, a set of initial themes is identified and may be written on the margins. The second reading is done and the themes are confirmed (and revised); also new themes may be added. This is done several times. Sometimes independent coders are used to check the credibility of deriving low-level themes from the text.
3. Low-level themes are grouped into a smaller number of high-level themes. This grouping involves an element of interpretation on the

part of the researcher: they need to decide if X, Y and Z belong to category A. As a credibility check, other researchers may be involved in the process so that results of grouping can be compared across researchers. The result of this stage of analysis is a manageable set of high-level meaningful units that summarize the transcript.

4. A summary table of themes is prepared. The table lists all the high-level emergent themes, all the lower-level themes within them, and supporting quotations from the raw transcript. The structure of themes can also be revised slightly at this point to account for parts of the transcript that are still unexplained. Data saturation is reached when subsequent readings of the transcript do not lead to identifying any new themes.
5. Finally, conclusions are formulated based on the summary table. These conclusions link the emergent themes to the theory. As a credibility check, participants may be shown the results of the analysis and asked to confirm the emergent themes as well as the derived interpretations.

The resulting analysis may be accompanied by "**memos**" that explain to the reader how and why certain analysis decisions were made, increasing the "thickness" of descriptions (which, as you know, increases credibility).

Inductive content analysis can also be applied to observational data. In this case the raw material for analysis comes in the form of field notes describing a participant's behaviour rather than interview transcripts.

If a theory emerges from the data, it is referred to as a "**grounded theory**". The name suggests that grounded theory "grows out of" empirical data as opposed to prior beliefs.

Exercise

- Find an example in this book of a study that used the interview or the focus group as the primary research method. What type of interview or focus group was it? How was content analysis organized?
- What can you say about generalizability and credibility of the findings?

Case study

A case study is an in-depth investigation of an individual or a group. You might say that this is not a proper definition because other research methods can also be defined this way, and you would be right. In fact, case studies can involve a variety of other methods (observations, interviews, and so on), anything that deepens our understanding of an individual or a group of interest. There are several reasons why case studies are referred to as a separate research method, even though they are actually a combination of other methods.

- The individual or group that is the object of a case study is unique in some way. As a result, the purpose is to gain a deep understanding of this particular individual or group.
- Sampling is not an issue: you are interested in this particular case, not the population this case "represents".
- There is less focus on generalizability. Findings do get generalized, but this is a by-product of the in-depth description and explanation of the case (case-to-case and theoretical generalization).
- The case is studied thoroughly, using a combination of different methods, and often longitudinally. This is why we defined a case study as an "in-depth investigation".

What are the reasons for choosing a case study as the preferred method?

First, case studies are useful to investigate phenomena that could not be studied otherwise. For example, it is a group that is hard to get access to and you may only get a chance to study one individual (think about studying the personality of a serial killer).

Second, case studies can contradict established theories and help develop new theories. Why is this a good thing? According to the principle of falsification in science (Karl Popper), the proper way to test a theory is to find one case that contradicts it. If you cannot, the theory stands, but if you succeed, the theory needs to be rejected or modified, and this is how science develops. To test the theory that "all swans are white" you need to try and find one black swan. In a similar fashion, universal theories of memory in cognitive

psychology can be tested by studying individuals with unusual or unique memory abilities. If in these individuals memory proves to function differently, then the universal theory of memory is not as universal as we thought. So, "boundary" cases are interesting, and since they are quite rare, we want to study them thoroughly.

Case studies have several limitations. Researcher bias can be a problem as, due to the longitudinal nature of the study, researchers might get too involved. Participant bias is also a potential problem for the same reason: the participant interacts with the researcher for a long period and it is easier for the participant to become susceptible to acquiescence, social desirability, and so on. The generalization of findings is especially problematic from a single case to other settings or to a wider population. Generalization depends on thickness of descriptions and triangulation (other researchers, other case studies, and so on).

Apart from the ethical considerations involved in qualitative research in general, case studies are especially demanding in terms of anonymity and confidentiality—it is difficult to preserve anonymity of unique cases. In case studies of patients with brain damage it may be difficult to obtain informed consent because they might not fully realize the terms of the document. It is debatable how "informed" this informed consent is exactly. In cases like this it is usually a parent or spouse who has overall responsibility for the patient and gives consent.

Ethics in psychological research

Inquiry questions

- Since psychology is a study of living beings, what ethical issues does it raise?
- How can we decide what is ethical and what is not in psychology?

What you will learn in this section

- Ethical considerations in conducting the study
  - Informed consent
  - Protection from harm
  - Anonymity and confidentiality
  - Withdrawal from participation
  - Deception
  - Debriefing
  - Cost-benefit analysis in ambiguous cases
- Ethics committees
- The Little Albert experiment
- Ethical considerations in reporting the results
  - Data fabrication
  - Plagiarism
  - Publication credit
  - Sharing research data for verification
  - Handling of sensitive personal information
  - Social implications of reporting scientific results
- The controversy around Cyril Burt

Ethics is an integral part of psychological research because it is research with living beings (humans and animals). This is one of the things that distinguishes the human sciences from the natural sciences—ethically, the study of human beings is not the same as the study of material objects.

All around the world the activities of psychologists are regulated by codes of ethics. These codes outline the ethical principles and procedures to be followed in all aspects of a psychologist's professional activities: counselling, testing and research. If a psychologist breaches the code, his or her professional license may be discontinued. Codes of ethics have been developed by international as well as national psychological associations, and there is a lot of overlap in their content as the ethical considerations in psychology are pretty much universal.

Exercise

Explore the Code of Ethics on the website of American Psychological Association (APA) and the Code of Human Research Ethics by British Psychological Society (BPS).

Compare the two codes and make a poster for your classroom highlighting the main similarities and differences:

APA:
http://www.apa.org/ethics/code/.

Exercise (continued)

BPS:
http://beta.bps.org.uk/news-and-policy/bps-code-human-research-ethics-2nd-edition-2014

Since IB psychology is an academic subject (involving no counselling), we will only focus on ethical considerations related to research. We will also break them into two large groups:

- ethical considerations in conducting the study
- ethical considerations in reporting the results.

Ethical considerations in conducting the study

The following list outlines the main ethical considerations to be addressed when conducting a research study in psychology.

- **Informed consent**. Participation in a study must be voluntary, and participants must fully understand the nature of their involvement, including the aims of the study, what tasks they will be exposed to and how the data will be used. Researchers should provide as much information as possible and in the clearest possible way, hence the name "informed" consent. If the participant is a minor, consent should be obtained from parents or legal guardians.
- **Protection from harm**. At all times during the study participants must be protected from physical and mental harm. This includes possible negative long-term consequences of participating in a research study.
- **Anonymity and confidentiality**. These two terms are often used interchangeably, but they refer to slightly different things. Participation in a research study is confidential if there is someone (for example, the researcher) who can connect the results of the study to the identity of a particular participant, but terms of the agreement prevent this person from sharing the data with anyone. So, the participant provides personal data, but the data stays confidential under the research agreement. Participation in a study is anonymous if no one can trace the results back to a participant's identity because no personal details have been provided. An example of anonymity would be filling out an online survey without providing your name.
- **Withdrawal from participation**. It must be made explicitly clear to participants that, since their participation is voluntary, they are free to withdraw from the study at any time they want. Researchers must not prevent participants from withdrawing or try to convince them to stay.
- **Deception**. In many cases the true aims of the study cannot be revealed to the participants because it would change their behaviour (for example, due to social desirability). So a degree of deception needs to be used. In some research methods deception is part of the process (for example, covert observation). Researchers must be careful and if deception is used, it must be kept to the necessary minimum.
- **Debriefing**. After the study participants must be fully informed about its nature, its true aims, how the data will be used and stored. They must be given an opportunity to review their results and withdraw the data if they want to. If deception was used, it must be revealed. Care must be taken to protect participants from any possible harm including long-term effects such as recurring uncomfortable thoughts. In some cases psychological help must be offered to monitor the psychological state of the participant for some time after the study (for example, in sleep deprivation studies).

ATL skills: Self-management

To memorize short lists, it is useful to use acrostics—phrases in which the first letter of each word stands for one of the elements on the list. For example, the ethical considerations in conducting a study may be combined in the following acrostic:

Can (consent)

Do (debriefing)

Cannot (confidentiality)

Do (deception)

ATL skills (continued)

With (withdrawal)

Participants (protection from harm)

Try making such acrostics of your own with other lists in this unit: threats to internal validity, types of bias in qualitative research, and so on.

Display the results in your classroom to share with others and gradually you will pick out the ones that are most easily memorized.

Very often ethical decisions prior to conducting a study are not easy, and a **cost-benefit analysis** needs to be conducted. For example, sometimes participants should not know the true aim of the study for their behaviour to be more natural. Sometimes it is difficult to preserve confidentiality (for example, in unique cases). Sometimes there is a risk that participants could get mentally or physically harmed. For example, in the famous Stanford Prison Experiment (Haney, Banks and Zimbardo, 1973) participants were led to believe that they were imprisoned and were kept in harsh conditions, being humiliated and dehumanized by other participants (who were randomly assigned the role of guards). Studies of such phenomena as obedience, conformity, compliance, violence and prejudice can rarely be designed so that they are harmless to the participants. So can we make the decision to relax some of the ethical standards for a particular study?

Such decisions can be made in some circumstances, including:

- if potentially the study can reveal scientific information that will benefit a lot of people
- if there is no way the study of a phenomenon can be conducted without relaxing an ethical standard.

In all countries professional bodies of psychologists have **ethics committees** that resolve ambiguous issues and approve research proposals. Research proposals with a full description of the aims, procedures and anticipated results are submitted to the committee and reviewed. In some cases, when research is potentially useful, ethically ambiguous research studies may get the "green light". Then the researchers will need to be extra careful in making sure that participant harm is minimized and long-term follow-up after the study is provided. Failure to cooperate with an ethics committee is itself a violation of ethics.

Psychology in real life

If you want to know more about the Stanford Prison Experiment, explore this website: http://www.prisonexp.org/.

You may also find Philip Zimbardo's TED Talk "The psychology of evil" interesting: https://www.ted.com/talks/philip_zimbardo_on_the_psychology_of_evil.

Research in focus: The Little Albert experiment

The Little Albert experiment was carried out by John B Watson (Watson and Rayner, 1920). The study provided evidence of classical conditioning in humans. Similar to Ivan Pavlov's experiments with his dogs (salivating at the sound of a bell), Watson was trying to form a certain reaction in response to a certain stimulus in a human baby. Watson observed that a baby's fearful reaction to loud noises was an innate, automatic response. When they hear a loud noise, little children always display behavioural signs of fear (tears, and so on). So he set out to form a fearful reaction to a neutral stimulus, furry objects, using the classic Pavlovian techniques.

Now he fears even Santa Claus

▲ Figure 1.12 Little Albert experiment

Research in focus (continued)

Their participant was a nine-month-old infant from a hospital who was referred to as "Albert" for the purposes of the experiment. During the baseline test Albert was exposed to a white rat, a rabbit, masks with hair, cotton, wool and other objects. Albert showed no fear in response to these objects. During the experiment a white laboratory rat was placed in front of Albert and he played with it. Every time the baby touched the rat, however, researchers hit a suspended steel bar behind his back with a hammer, producing a very loud sound. Naturally, the baby cried and showed fear. After pairing these two stimuli several times, the steel bar was taken away and Albert was only presented with the rat. In line with the Pavlovian theory, Albert would show signs of distress, cry and crawl away. So, the researcher "succeeded" in forming a fear of a rat in a baby. In further trials it was revealed that fear in Little Albert was actually generalized to other furry objects. He would show distress, cry and crawl away at the sight of a rabbit, a furry dog and even a Santa Claus mask with a beard.

As you can see, the study exposed the infant to severe distress and potential long-term detrimental consequences. To make things worse, Albert left the hospital (taken away by his mother who did not leave any contact details) shortly after the experiment, and although Watson had planned to carry out de-sensitization, he never had the opportunity. So Albert returned to his daily life with a set of newly formed phobias, and without ever realizing why he had them.

What are the major ethical issues in this study? How would you go about conducting the study in a more ethically appropriate way?

Ethical considerations in reporting the results

The following list gives the main ethical considerations to be addressed when reporting results.

- **Data fabrication.** This is a serious violation of ethical standards and psychologists may lose their license if they fabricate data. If an error is found in already published results, reasonable measures should be taken to correct it (for example, retraction of an article or publication of an erratum).
- **Plagiarism.** It is unethical to present parts of another's work or data as one's own.
- **Publication credit.** Authorship on a publication should accurately reflect the relative contributions of all the authors. For example, the APA Code of Ethics states specifically that if a publication is based primarily on a student's work, the student must be listed as the first author, even though his or her professors co-authored the publication.
- **Sharing research data for verification.** Researchers should not withhold the data used to derive conclusions presented in the publication. The journey from raw data (in the form of a matrix with numbers for quantitative research or a text/transcript for qualitative research) to inferences and conclusions is full of intermediate decisions, interpretations and inevitable omissions. It is healthy scientific curiosity to want to replicate the analysis, and any request from an independent researcher to share raw data should be satisfied, provided both parties use the data ethically and responsibly. This entails, for example, making the shared data set anonymous (deleting the names or other identifiers) and only using the shared data set for the stated purposes.
- **Handling of sensitive personal information.** This refers to how the results of the study are conveyed to individual participants.
  - **Handling of information obtained in genetic research.** Research into genetic influences on human behaviour, such as twin, adoption or family studies, can sometimes lead to revealing private information to one individual about other members of the person's family. Examples include misattributed parentage or health status. In twin studies one may discover that he

or she has a twin that he or she has never met. Information of this sort may be disclosed accidentally during interviews, inferred by the participants in the debriefing session or in the report of results. All these considerations imply certain requirements in the way results should be relayed to participants. Such information must be handled with care and sensitivity, and if detrimental consequences are suspected, subjects should be monitored for some time after the end of the study, and psychological counselling may be offered.

- **Handling of information related to mental disorders.** Some studies may result in revealing the presence of illness that was previously unknown (for example, a study of depressive symptoms in response to life stress requires carrying out a diagnosis of depression for all participants). This knowledge may have a lot of unwelcome consequences such as a change in self-esteem or a change in family perceptions and expectations for a child. On the other hand, research may reveal that some family members do not have the disease now, but they are at higher risk of developing it in the future. People may not want to know that.
- **Social implications of reporting scientific results.** Researchers must keep in mind potential effects of the way research conclusions are formulated on the scientific community and society in general. For example, imagine you conducted a research study that supported the idea that homosexuality is inherited. Where should you publish the results? Should it be a narrowly specialized scientific journal or a more popular journal that targets a wider audience including non-scientists? Stating that homosexuality is inherited (and bluntly believing in this statement because it "came from the scientists") may have deep effects on society. At the same time, you can never be sure of the results of a single research study—there might have been bias; measurements might have been inaccurate; findings may later turn out to be false. Science is a very meticulous (and often inconclusive) process, and care must be taken to report results precisely and accurately, recognizing all potential limitations of the research study, especially if the findings are of social significance.

Research in focus: The controversy around Cyril Burt

There is much controversy about the work of Cyril Burt, a British psychologist who became famous for his contributions to intelligence testing. In 1942 he became president of the British Psychological Society. He was responsible for administration and interpretation of mental ability tests in London schools. In one of his most famous studies he conducted research with 42 identical twins reared apart. His results showed that the IQ scores of identical twins reared apart were much more similar than that of non-identical twins reared together. He concluded that genetic inheritance in intelligence plays a much greater role than environmental factors (such as education).

In 1956 Burt reported on another study, this time with 53 pairs of identical twins raised apart, where he found a high correlation (0.771) between the IQ scores of the twins. This was exactly the same correlation (to the third decimal place) that he had reported in an earlier study with a smaller sample size. Burt's research was very influential in forming educational policies in the country, for example, the belief that intelligence is fixed and hereditary led to the practice of using standardized tests to measure intelligence in school children and allocate them to schools based on the results.

After his death in 1971 the British Psychological Society found him guilty of publishing a series of fraudulent articles and fabricating data to support the theory that intelligence is inherited. The case was built on several details that were considered to be highly suspicious.

- There was a very unlikely coincidence of the same correlation coefficient in the two studies.

Research in focus (continued)

- Some factors that should theoretically influence intelligence (such as mental illness or childhood influences) were suspiciously unimportant in Burt's data sets, almost a statistical impossibility.
- Identical twins reared apart is an extremely rare sample; there were only three other studies at that time using this kind of sample and none of them had more than 20 pairs of twins as participants.
- Burt's two female collaborators who worked for him collecting and processing data could not be found, their contact with Burt could not be traced and it was even suspected that these people never existed!

However, some scholars have recently re-examined the claims made earlier and found that evidence of Burt's fraud is not conclusive, or at least he deserved the benefit of doubt.

▲ Figure 1.13 Cyril Burt

In any case, data sets and publications that raise questions regarding their credibility are in themselves an ethical concern, even if they are not falsified intentionally. This is especially true for settings where research findings are used to inform social (for example, educational) policies.

Exercise

At the beginning of this unit you came up with a research proposal related to a research question. Go back and review that proposal. Now that you are equipped with more knowledge about research methodology in psychology, what would you change in your original proposal and why?