Module 1: Origins

1.01 Origins: Non-scientific methods

To see why we need the scientific method, let's take a look at what people base their knowledge on in day-to-day life. People can accept something as true based on **intuition** or **belief**. Let's consider my own strong belief that my cat Misha loves *me* most of all people in his life. I just *know* he loves me more than anyone else, I feel this in my heart of hearts.

Is such a belief a good basis for knowledge? Well no, simply believing something doesn't make it so. Things we believe in strongly can turn out to be false. Also, what if someone else holds an opposing belief? What if my fiancé believes that Misha loves *him* more? There is no way to settle who is right *just* by pitting our beliefs against each other.

We could count the number of supporters for each belief and require a majority or **consensus**. But this isn't a very solid basis for knowledge either. Just because most people accept something as true doesn't mean it *is* true. For centuries practically everybody thought the earth was flat. Turns out they were wrong; it's round.

Another source of knowledge is an **authority**'s opinion; also not a very good source. The opinion of authority figures like political leaders, experts, scientists, is just that, an opinion. Authorities may have access to more or better knowledge but they also have an important personal stake in getting their views accepted. Their careers and reputation depend on it.

Suppose my fiancé gets a so-called cat-whisperer to declare that Misha loves him more. Of course I'm going to be skeptical about this expert opinion, especially if my fiancé paid for it. I could find my own cat expert to oppose my fiancé's cat whisperer but then we would just have two opposing opinions again. What we need is **evidence**.

So how do we use evidence to settle the argument of whom Misha loves more? Well, suppose I regularly *observe* that after getting home from work, Misha always comes to sit on *my* lap and not my fiancé's. I'm supporting my *statement about the world*, that Misha loves me more, with an *observation of the world*, namely on whose lap he sits after work.

This gathering of evidence through **casual observation** is a better foundation of knowledge than the previous ones, but still not good enough. This is because people just aren't very good at observing. We tend to selectively observe and remember things that agree with our beliefs. For example, I might have forgotten - very conveniently - that Misha always sits on my fiancé's lap at breakfast. There are many biases besides selective perception that make casual observation a tricky source of knowledge.

The same goes for our ability to use logic. Logical reasoning would seem like a solid basis for knowledge. But our informal logical reasoning isn't always consistent. There's an almost endless list of 'fallacies' or logical inconsistencies that people regularly make in their day-to-day reasoning.

If we want to develop accurate knowledge, make sure that our explanations of the world are valid, then we need something more. We cannot depend on subjective and unverifiable sources like beliefs, opinions and consensus; and we can't trust casual observation and informal logic because they can be heavily distorted by our beliefs. We need ***systematic* observation**, free from any bias, combined with **consistently applied logic**. In other words, we need the **scientific method**.

1.02 Origins: Scientific method

We need the **scientific method** to make sure our attempts to explain how the world works result in valid knowledge. Opinions, beliefs, casual observation and informal logic won't do; they are too *subjective* and too susceptible to *error*.

The scientific method is based on *systematic observation* and *consistent logic*. Applying the scientific method increases our chances of coming up with valid explanations. It also provides a way to evaluate the plausibility of our scientific claims or *hypotheses*, and the strength of the empirical evidence that we provide for these hypotheses in our empirical study or research.

The scientific method can be described according to six principles. If our study meets these principles, then it can be considered scientific. Our hypothesis can then be compared to, and compete with other scientific claims to provide the best possible explanation of the world around us.

The first principle requires that a hypothesis is **empirically testable.** This means that it should be possible to collect empirical or physical evidence, or *observations*, that will either support or contradict the hypothesis.
Suppose I hypothesize that my cat loves me more than he loves my fiancé. To test this hypothesis empirically we need to collect observations, or *data*. But how can we observe how much the cat loves us? We can't ask the cat about his feelings. Suppose we both agree that a cat is unable to express love the way humans do. Well, then there is nothing to observe; the hypothesis is not empirically testable.

The second principle is **replicability**. A study and its findings should be **replicable**, meaning we should be able to consistently repeat the original study. If the expected result occurs only *once* or in *very few* cases, then the result could just have been coincidental. A hypothesis is more plausible if it is repeatedly confirmed. And this requires that it is possible to repeat or replicate a study.

Let's say I've convinced my fiancé that if the cat loves someone more, the cat will spend more time on their lap. Now suppose I observed that this week the cat sat on my lap twice as long as on my fiancé's lap. Does that mean my hypothesis can be accepted? Does the cat love me more? Well, the hypothesis would be considered plausible if we can show that the result is the same in the following weeks. But what if the cat dies after the first week of observation? Then we would not be able to check the hypothesis for ourselves. The study is no longer replicable!

To see if results replicate, we have to be able to repeat the study *as it was originally conducted*. Suppose we do something differently and we find different results. Is this a failure to replicate? No, the failed replication could be caused by our change in procedure.

The third principle of Objectivity aims to allow others to repeat the study by themselves, without need for the original researcher. **Objective** literally means that it shouldn't matter who is performing the study.
Anybody should be able to get the same results based on the description of the assumptions and procedures. A researcher should therefore be as **objective** as possible about assumptions, concepts and procedures. This means that all these elements should be *clearly* and *explicitly defined*, leaving no room for subjective interpretation.

Suppose I count my cat's face rubbing as an expression of love, but I fail to explicitly tell my fiancé about this. Then my procedure for measuring love is subjective. Even if we systematically observe the cat at the same time, the result will depend on who is observing him. I will conclude the cat shows love more often than my fiancé will.
In this example, the results are subjective and therefore incomparable, and we might not even be aware of it. If we do not explicitly discuss and agree on what counts as love and what doesn't, then our measurement procedure for cat love is not objectively defined.

The fourth principle is **transparency**. Being **transparent** is closely related to being objective. In science *anyone* should be able to replicate your results for themselves, your supporters but also your critics. This means that researchers need to ***publicly share*** what assumptions were made, how concepts are defined, what procedures were used and any other information that's relevant for accurate replication.

The fifth principle states that a hypothesis should be **falsifiable**. Falsifiability is a very important principle. A hypothesis is falsifiable if we are able to at least *imagine* finding observations that will contradict our hypothesis. If we can't imagine what such contradictory data would look like, well then the hypothesis cannot be disproven.

Ask any person with a very strong, for example, religious belief what evidence would convince them that their belief is false. No matter what contradictory evidence you propose, they will probably argue that these facts do *not* contradict their strong belief. This puts statements based purely on belief, such as religion, outside the domain of science. If there is no form of evidence that will be accepted as disproving a hypothesis, then it is pointless to argue about the hypothesis or

to even look for confirmation, since the conclusion is already drawn.

Ok, let's move on to the sixth and last principle of **logical consistency**. A hypothesis should be logically consistent or coherent. This means there shouldn't be any internal contradiction, for example if a supporting assumption disagrees with the hypothesis. The conclusions based on our observations should also be *logically consistent*. This means, among other things, that researchers should be consistent in what they count as confirmatory and contradictory evidence.

Let me explain this using our cat example: I hypothesized that my cat loves me more and so I expect him to sit on my lap longer. What if he spends more time on my fiancé's lap? I could say that the cat can feel that sitting on my lap is uncomfortable for me. So the cat will sit on my lap less often *because* he loves me more. Of course this is logically inconsistent. I've changed the interpretation of the results after the data are in to suit my hypothesis. Incidentally, this also makes my hypothesis unfalsifiable; I will always conclude that my cat loves me, whether he sits on my lap often or not at all.

So to summarize, the scientific method requires that we formulate hypotheses that are:
empirically testable: meaning the hypothesis can be supported or contradicted by *observations*;
replicable: meaning the hypothesis can be tested *repeatedly*;
objective: meaning the hypothesis can be tested *independently* by others;
transparent: meaning the hypothesis and results are *publicly shared* so they can be tested *by anyone*;
falsifiable: meaning that finding *contradictory evidence is a possibility*;
and finally: **logically consistent**: meaning that the hypothesis is internally consistent and the conclusion to support or reject the hypothesis, based on the observations, is logically sound.

One final point: the scientific method is only effective when it is used with the right attitude. In order to come up with better hypotheses, researchers need to be critical of their own studies and those of others. This means they have to be **open** and **transparent**; they have to accept critique and let go of their pet-hypotheses if others provide better explanations. Only then can science function like an evolutionary system, where only the fittest, or most plausible hypotheses survive.

1.03 Origins: Scientific claims

Until now I've talked about *statements*, *hypotheses* and '*explanations* of the world around us'. And I've used these general terms without specifying what they mean exactly. It's time to clarify this.

Scientific claims about the world around us can be categorized into different types. Some scientific claims describe or explain more phenomena than other claims. Also, some scientific claims provide more plausible descriptions or explanations of the world around us. We find some claims to be more certain, better supported by evidence, than others. In *science* the most basic claim is an **observation**. An observation can be an accurate or inaccurate representation of the world. Suppose I observe that my cat, which has a ginger-colored coat weighs 6.5 kilograms.

Most scientists would accept this observation as a probably fairly accurate reflection of a specific aspect of the world around us, assuming the weight scale is valid and reliable. But in terms of explanatory power, they would find this observation very uninteresting, because an observation on it's own is not very informative; it doesn't describe a ***general relation between properties*** and it doesn't ***explain*** anything.

That doesn't mean observations are unimportant; Observations are the *building blocks* of the empirical sciences. But they're not very useful on their own. An observation on its own is the least interesting type of scientific claim since it has no explanatory power. Observations become useful when they are used to confirm or contradict a **hypothesis**. A hypothesis is a statement that describes a *pattern* or *general relation* between *properties*. A hypothesis can also *explain* the pattern that it describes.

Take this hypothesis: ginger cats will on average be overweight more often than cats with a different color fur. And I could extend this hypothesis with an *explanation* for the relation between fur color and obesity, for example by stating that the genes for ginger fur color and signaling fullness of the stomach are linked.
The plausibility of a *hypothesis* can range from very uncertain to very certain. A hypothesis can be unsupported and therefore uncertain, for example if it's new and still untested. A hypothesis can also be strongly supported by many empirical studies and therefore more certain.

A special type of hypothesis is a **law**. Laws are very precise descriptions of relations or patterns; so precise that they are usually expressed as mathematical equations. They are also generally very well-substantiated; that's how they got so precise.
For example if I drop my cat's food bowl from a height of 56 meters and I know the earth's gravitational constant, then I can predict very accurately how long it will take for the bowl to hit the ground, by using Newton's gravitational laws. Laws allow for very precise predictions, but they usually don't *explain* the relationships they describe, in this case between distance, time and gravity.

Of course in the social sciences laws are hardly ever formulated. We understand too little of people and groups yet to be able to specify patterns in their behavior with such a degree of precision that we can postulate scientific laws.

Ok, so this leaves us with the term **theory**. In day-to-day life 'theory' means an unsubstantiated statement, an educated guess. In *science* however, 'theory' refers to a broad, overarching explanation of many related phenomena. In the natural and behavioral sciences, a theory is built up out of hypotheses that are *very strongly supported by empirical evidence*.

In the social sciences, where *qualitative* and *historical comparative approaches* are more dominant, a theory is considered highly plausible when it has withstood attempts to refute it, based on logical grounds as well as historical or qualitative analysis. So in science, theories are the most well-established explanations, the closest thing to certainty that we have, because they consist of hypotheses that have survived the scrutiny of the scientific method.

Of course this doesn't mean that scientific theories are certain or true. There have been many well-substantiated theories that were ultimately replaced, like Newton's mechanics that made way for the special theory of relativity. In science there is no certainty, only a provisional best explanation.

1.04 Origins: Classical period

The first thinkers to seek natural or earthly explanations instead of divine explanations were ancient Greek scholars like Thales, Pythagoras and Democritus. But the first to really consider *how* to obtain knowledge were Plato and Aristotle, more than 2.300 years ago.

To **Plato** the external world and the objects in it are just imperfect reflections, or shadows, of *'ideal'* forms. These ideal forms are often portrayed as casting shadows on a wall. Plato was a philosophical *realist*; he thought reality, in his case the world of forms, exists *independently* of human thought. To Plato these forms are not just abstract concepts in our mind, they really exist, but separately from the physical world.
Plato thought that since the physical world we see is an imperfect reflection of reality, we can't learn the true nature of reality through sensory experience. He insisted that knowledge about the ideal forms can only be gained through *reasoning*. Plato is therefore referred to as a *rationalist*.

Plato's student **Aristotle** was a *realist*, just like Plato. He thought that reality exists independently of human thought. But to Aristotle reality *is* the physical world. There is no separate plane of existence where abstract forms live.
Aristotle also disagreed with Plato on how we can gain knowledge about the true nature of things. Aristotle was an *empiricist*. He believed our

sensory experience gives an accurate representation of reality, so we can use our senses to understand it. He believed that ultimately, knowledge comes through *observation*.

But that doesn't mean Aristotle was interested in observations only. He still saw reasoning as the best way to understand and *explain* nature; he in fact developed **formal logic**, more specifically the *syllogism*. Here's an example of a syllogism:
"All humans are mortal, all Greeks are humans, and therefore all Greeks are mortal". If the two premises are true, then the conclusion is necessarily true. By using this conclusion as a premise in a new syllogism, our knowledge builds.

Of course this only works if the premises are true. Consider this one: "All mammals are furry, all cats are mammals, therefore all cats are furry". The first premise is false, which means the conclusion is not necessarily true. Not a good basis for building knowledge! So how can you be sure a premise is true? Well you can prove it using another syllogism, but of course you have to keep proving *those* premises, so there has to be a set of starting premises that you can accept as undisputedly true.

According to Aristotle these **fundamental premises** can be determined through *observation* of basic patterns or regularities in the world. Unfortunately he wasn't aware that some of his own observations were too selective, leading to fundamental premises that we know now are just plain wrong. For example, he thought, based on his observations, that insects have four legs, and that men have more teeth than women.

Aristotle probably came to these conclusions based on observations of the mayfly which walks on four legs, but like other insects actually has six legs; it's also likely that he examined his own teeth and those of male friends but only examined the teeth of servant-women who were more likely to be malnourished and have less teeth. He didn't realize it, but his observations were inaccurate. Even so, Plato's and Aristotle's views remained dominant for almost 2000 years! It took until the end of the 16th century for people to realize that Plato and Aristotle's views were flawed.

How did the scientific method develop after Plato and Aristotle? Well, the ancient Greeks made many scientific advances. For example, **Ptolemy** described the movement of planets by placing the earth at the static center of the universe with the planets, including the sun, in a circular orbit, each moving in their *own* little cycle along their orbital path.
These cycles within cycles were necessary to explain the weird phenomenon of retrograde motion, where planets would sometimes move backwards. Ptolemy's model allowed for accurate predictions, but it's thought that people didn't really believe that it described the *actual* motion of the planets; it only *'saved the phenomena'*.

After the demise of the Greek city-states, during the rise and fall of the Roman Empire and the first centuries of the middle ages, very few scientific advances were made. Plato's and later Aristotle's philosophical ideas remained dominant until a new scientific revolution at the end of the 16th century, starting the age of enlightenment.

But, let's look at the developments that led up to that revolution. First, around the turn of the 10th century, Arab and Persian scholars such as **Ibn al-Hasan**, **Al Biruni** and **Ibn Sina** started using *systematic observation* and *experimentation*, emphasizing *unbiased observation* and not just logical reasoning.
Second, building on the work of their predecessors, the Englishmen **Grosseteste** and **Roger Bacon** advocated the use of both *induction* and *deduction*. Induction means using particular observations to generate general explanations. Deduction means predicting particular outcomes based on general explanations.
A third important development was the invention of the printing press. This created the perfect conditions for a scientific revolution. More scholarly works became available to a wider audience. Among these works was "*De revolutionibus orbium coelestium*" by **Copernicus**.
This was the fourth important development to lead up to the scientific revolution. In Copernicus' new model of planetary motion, the planets, including earth, moved in circles around the *sun*.

Now this didn't exactly agree with religious doctrine; the Church accepted Aristotle and Ptolemy's model with *earth* at the center of the universe. Many historians believe Copernicus was afraid to publish his work because he feared the Church would punish him for contradicting their doctrine. He did eventually publish his new model. But he added a special dedication to the pope, arguing that if *Ptolemy* was allowed to formulate a model with strange cycles that only *'saved the phenomena'*, well then he should be given the same freedom.

He was implying that *his* model was also intended, not as an accurate representation, but just as a pragmatic model. Whether he truly believed this is unclear. He died shortly after the publication, which actually did not cause an uproar until 60 years later.
Now according to many the scientific revolution and the age of enlightenment started with Copernicus. But others feel the honor should go the first man to refuse to bow to the Catholic Church and maintain that the heliocentric model actually described physical reality. This man of course, was **Galileo Galilei**.

1.05 Origins: Enlightenment

Galileo is considered the father of modern science because he set in motion the separation of science from philosophy, ethics and theology, which were all under strict control of the Catholic Church. Others had already quietly advocated a scientific approach based on observation and experimentation, instead of using theological reasoning. But Galileo was the first to do this very explicitly.

He also opposed several of Aristotle's theories, which were accepted by the Catholic Church as doctrine. For example, he disproved the Aristotelian view that heavy objects fall to the earth more quickly than lighter objects. Galileo did

this with a thought experiment, showing that besides observation, he also valued logical reasoning.

Of course he is most famous for disputing the Aristotelian and Ptolemaic view that the earth is the center of the universe. He supported Copernicus' heliocentric view, where the sun is the center of the universe. Galileo made systematic observations of the planet Venus that could only be explained if the planets revolved around the sun, instead of earth.
Now, to Copernicus the heliocentric model just '*saved the phenomena*', meaning that the model accurately predicts our observations of planets, but that it doesn't actually correspond to physical reality. In contrast, Galileo had no problem claiming that the earth really revolves around the sun. The Catholic Church did not appreciate Galileo's disruptive ideas. They brought him before the inquisition and put him under house arrest until his death.

René Descartes, of the Cartesian coordinate system, was a contemporary of Galileo. Although Descartes also rejected many of Aristotle's ideas, Descartes did agree with Aristotle that knowledge should be based on first principles. Because he felt our senses and mind can be easily deceived, he decided to discard every notion that is even the least bit susceptible to doubt. And once he had removed everything that he doubted, he was left with only one certainty, namely that he thought and therefore he must exist. '*Cogito ergo sum*'. This eventually led him to conclude that we only know the true nature of the world through reasoning.

Francis Bacon thought, just like Descartes, that scientific knowledge should be based on first principles. But in contrast to Descartes, Bacon maintained that this should happen through *inductive methods*. Induction means that observations of particular instances are used to generate general rules or explanations. Suppose every time I've encountered a swan, the swan was white. I can now induce the general rule that *all* swans are white.
Bacon believed that all knowledge, not just the first principles, should be obtained only through this inductive method, generating explanations based on sensory experiences. This is why he is considered the father of **empiricism**, where empiric means relating to experience or observation.

Now, David **Hume** took empiricism to the extreme, accepting *only* sensory data as a source of knowledge and disqualifying theoretical concepts that didn't correspond to directly observable things. This led him to conclude that the true nature of reality consists only of the features of objects, not of the physical objects themselves. This extreme form of empiricism is called **skepticism**.
I'll give you an example. Let's take as a physical object a cat. Now what makes this cat a cat? Well its properties: its tail, whiskers, coloring, fur and body shape... If you take away all the properties that make it a cat you are left with… well, nothing. The essence of the cat is in its features.

Hume also showed us the **problem of induction**: even though you've consistently observed a phenomenon again and again, there is no guarantee

your next observation will agree with the previous ones. For a long time, from the perspective of Europeans at least, all recorded sightings of swans showed that swans are white. Only after Australia was discovered did we find out that there are also black swans.
In other words, no amount of confirmatory observation can ever conclusively show that a scientific statement about the world is true. So if you require that all knowledge must be based on observations alone, that means you can never be sure you know anything!

Partly in reaction to Hume's skepticism, at the start of the 19th century a philosophical movement known as **German idealism** gained popularity. The idealists believed that we mentally construct reality. Our experience of the world is a mental reconstruction. Scientific inquiry should therefore focus on what we can know through our own reasoning. Now, the Idealists concerned themselves mainly with questions about immaterial things like the self, god, substance, existence, causality. They were also criticized for using obscure and overly complicated language.

On the eve of the Second Industrial Revolution around the turn of the 19th century, scientists started to lose patience with the metaphysics of the idealists. Their musings on the nature of being had less and less relevance in a period where scientific, medical and technical advances were rapidly being made.
At the start of the 20th century a new philosophy of science, came on the scene that proposed a radical swing back to empiricism. This movement is called logical positivism.

1.06 Origins: Modern science

After the First World War, a group of mathematicians, scientists and philosophers formed the *Wiener Kreis*, in English called the *Vienna circle*. They were unhappy with the metaphysics of the German idealists, who focused on first principles of knowledge and the fundamental nature of being.
The Vienna circle, with members like Moritz *Schlick*, Otto *Neurath* and Rudolf *Carnap*, felt idealist questions about the self and existence were meaningless because they were unanswerable. They proposed a new philosophy of science called **logical positivism**.
The logical positivists redefined science as the study of **meaningful statements** about the world. For a statement to be meaningful it has to be verifiable, which is known as the **verification criterion**. It means that it should be possible to determine the truth of a statement.

There are two types of meaningful statements. Analytic statements and synthetic statements. **Analytic statements** are tautological, necessarily true. Examples are "bachelors are unmarried" and "all squares have four sides". They are **a priori statements**, like definitions and purely logical statements.
They don't depend on the state of the world and therefore don't require observation to be verified. They can be used in mathematics and logic. New

combinations of analytic statements can be verified with formal logic.
Synthetic statements depend on the state of the world. Examples of synthetic statements are: “All bachelors are happy” and: “All cats are born with tails”. These statements are **a posteriori**; they can *only* be verified through observation. The logical positivists thought these statements should be always **publicly accessible**.
Also, statements are not allowed to refer to **unobservable entities** like “electron” or “gravity”, because they can't be observed directly. If a statement makes reference to an unobservable entity, is not tautological or not logically or empirically verifiable, then that statement is meaningless. Subjects like metaphysics, theology and ethics were thereby nicely excluded from science.

Of course the criterion of verification through observation couldn’t deal with the problem of *induction*. No amount of confirmatory evidence is ever enough to definitively prove or verify a statement. It’s always possible a contradictory observation will be found in the future. So the strong criterion of verification was weakened by requiring only **confirmation** instead of verification.
Another very strict rule also had to be changed. Not allowing reference to unobservable entities created big problems. Entities like “electron”, ”gravity” and “depression” cannot be observed directly, but they are indispensable in scientific explanations. This, together with the problem of induction, led to a more moderate version of logical positivism called **logical empiricism**.

Karl Popper, who was nicknamed "the official opposition" by the Vienna circle, was one of their main critics. He argued that the distinction between meaningful and meaningless statements should be based on the criterion of ***falsification***, not verification.

Karl Popper argued that we can never conclusively verify or prove a statement with observations, but we can conclusively disprove it with contradictory evidence. According to Popper a statement is meaningful only if it’s **falsifiable**.
Popper proposes that scientists should actively engage in "risky experiments". These are experiments that maximize the chance of finding evidence that contradicts our hypothesis. If we find such contradictory evidence, we inspect it for clues how to improve our hypothesis. The hypothesis is provisionally supported, only if contradictory evidence is absent.

Now, Willard van Orman Quine showed that this criterion is also problematic. In the Duhem-Quine thesis, he states that no hypothesis can be tested in isolation; there are always background assumptions and supporting hypotheses. Now if contradictory evidence is found then according to Popper, our scientific explanation is wrong and should be rejected it. But according to Quine we can always reject one of the background assumptions or supporting hypotheses instead. This way we can salvage the original hypothesis.

Thomas Kuhn pointed out that science doesn't develop out of strict application of either the verification or the falsification principle. Hypotheses aren't immediately rejected or revised if the data don't agree with them. Science takes place

within a certain framework or paradigm. Hypotheses are generated that fit within this paradigm. Unexpected results lead to revision of hypotheses but only as long as they fit the framework. If this is impossible, the results are just ignored. But when more contradictory evidence accumulates, a crisis occurs, which leads to a paradigm shift. A new paradigm is adopted and the cycle begins again.

Even in it's weaker form of logical empiricism, logical positivism couldn't stand up to the critique of Popper, Quine and others. Since then, we've progressed to a more pragmatic philosophy of science. Today scientists follow the **hypothetico-deductive** method, combining induction and deduction, requiring falsifiability and accepting repeated confirmation only as provisional support for a hypothesis.

Philosophically, many scientists would probably be comfortable with **Bas van Fraassen's *constructive empiricism***, which states that science aims to produce empirically adequate theories. Knowledge requires observation, but unobservable entities are allowed. Accepting a scientific theory doesn't mean accepting it as definitive, a true representation of the world. According to a constructive empiricist, a scientific statement is accepted as true *as far as our observations go*; whether the statement truthfully represents the unobservable entities simply cannot be determined. We just have a current best explanation for our observations. That's it.

1.07 Origins: Epistemology

Before you accept the hypothetico-deductive method as the best way to gain knowledge about the world, there are at least two important philosophical questions *about knowledge* that you should answer for yourself.
The first question concerns the nature of reality: What is real, what exists, and therefore what is out there that we can gain *knowledge* of in the first place? The philosophical field that deals with these types of problems is called **ontology**: The study of being.
The second question concerns the way in which knowledge can be **acquired**. Assuming there is a reality out there that is in principle knowable, then what knowledge of reality is accessible to us and how do we access it?

The field of philosophy that is concerned with these types of problems is called **epistemology**, the study or theory of knowledge. I'll start with the last questions first. Assuming there is a reality out there that is knowable, how do we obtain this knowledge? Well there are many different *epistemological* views; I'll just discuss the two most important views here.

First there's **rationalism**. Rationalists hold that knowledge is gained through reason. Using our mind's capability for logical, rational thought, we can deduce truths about the world without having to resort to experience. Philosophers like *Plato* and *Descartes* coupled rationalism with the idea that at least some of the abstract concepts about the structure of nature are *innate, we* were born with them.

That means our mind simply has the capability of understanding these concepts because we already know them. We just have to "remember" or "recognize" them by using our reason.

Empiricism opposes this view. According to the empiricist view, sensory experience is the most important way, according to some strict empiricists even the only way, to obtain knowledge about the world.
Aristotle is considered the first empiricist. He thought that the foundational truths about nature come from sensory experience. We can obtain more knowledge through deductive reasoning, but observation is the basis of all our knowledge.

Aristotle didn't believe in innate ideas, in fact he coined the term "***tabula rasa***" to indicate everyone is born as blank slate: our knowledge is not predefined, the mind is open to any idea. Of course Aristotle wasn't a radical empiricist. He didn't object to rational thought entering into the mix and he wasn't worried about using abstract, not directly observable concepts.

I guess *Galileo* can be considered a moderate empiricist. He put a lot of emphasis on observation and experimentation but he also relied heavily on logical reasoning. Galileo in fact famously said that the book of nature is written in the language of mathematics. He had no problem using thought experiments and included references to "unobservables" in his hypotheses.

Later empiricist such as *Bacon*, but especially *Hume* and the *logical positivists* were very strict empiricists, maintaining that *only* sensory experience could lead to true knowledge about the world. They considered statements about unobservable, universal properties that cannot be observed directly, to be meaningless.

The contemporary flavor of empiricism, is *Van Fraassen*'s **constructive empiricism**. It emphasizes the role of sensory experience in both inductive and deductive methods, but it allows for theoretical terms that don't have physical, directly observable counterparts. In constructive empiricism, the aim is to come up with empirically adequate explanations, which can be considered 'true' – they accurately describe the world - as far as the observables go.
A constructive empiricist would say that the truth or falsity as far as the *un*observables go, simply cannot be determined. This recognizes that knowledge is provisional because it always remains possible that new contradictory evidence will be found someday.

1.08 Origins: Ontology

Let's turn to the subject of **ontology**, or the study of being, which asks: What is the nature of reality? Well, there are many competing views. And before we dive into the philosophical views themselves I'll first explain two main points on which these views differ from each other.

The first main point is whether reality exists *independently* of human thought.
When we refer to objects we perceive in the world, are we referring to *actual entities that exist outside of us*, or are we referring to mental representations that are constructed by our mind and that can only be said to *exist in our mind*?
The second main point concerns the ontological status of **particulars** and **universals**. With particulars I mean specific *instances* or occurrences in which a property can be observed. With universals, or *unobservables*, I mean general properties that cannot be observed directly.

Let me give an example. Love is a general property that we cannot observe directly, but that is instantiated, or expressed, in behavior. So when my cat climbs on my lap and takes a nap, that could be a *particular* instance of the *universal* property 'love'. Another example of an unobservable, *universal* property is gravity. Gravity is expressed in *particular* instances, for example when I drop my cat's food bowl and it falls to the ground.
So let's look at some different ontological views and see where they stand on the question of particulars versus universals and the question whether reality exists externally or only in the mind.

Idealism is a philosophical view that states that reality, as we perceive it, exists entirely in our mind. The existence of an external, physical world is irrelevant, since our perception of it is determined by our mental processes. Reality *is* in effect a mental construct. 'Gravity' and 'Love' exist, but only in our mind. The same goes for their particular occurrences. So an Idealist would say that the cat sleeping on my lap and the bowl falling to the ground are also mental constructions.
The question whether universal, unobservable entities are real, external, independent entities is therefore less relevant for Idealism because both particulars and universals are considered to exist; they're just both mental representations.

Idealism can be contrasted with **materialism.** Materialism is a position that accepts an external world independent of our mind. Materialism also states that everything in this independent physical reality consists entirely of matter. This means that everything is a result of the interaction of physical stuff, including our consciousness, feelings and thoughts. These are all byproducts of our brain interacting with the physical world. The exact opposite of idealism, it's material versus mental! Materialism is only about what stuff is made of. Like idealism, it's not strongly associated with a view on the distinction between universals and particulars.

Realism is a different position. Just like materialists, realists maintain that external reality exists, independent of human thought. But realists also maintain that universals like Love and Gravity are 'real'. In what form these exist depends on your flavor of realism. **Platonic realism** refers to Plato's position that universals like Gravity and Love really exist independently from our observation, but on a separate, abstract plane.

Scientific realism is more moderate and states that it's possible to make consistently supported claims using universals in statements about observable phenomena. In scientific realism, universals like Love and gravity are therefore given the same ontological status as observable particulars. Unobservables are assumed to exist, since they're useful and often even necessary to formulate successful scientific claims.

Finally we have **nominalism.** This view opposes realism as far as universals are concerned; it accepts reality as independent of human thought but denies the existence of universals. In nominalism there is no such thing as Gravity or Love; there are only falling objects and cats that frequently sit in your lap purring. According to nominalists, we just use the terms Gravity and Love because they help us to make sense of the world, but these universals don't actually exist.

1.09 Origins: Approaches

The development of the scientific method I've discussed up until now was focused mainly on the *natural sciences*: physics, astronomy, biology. But during the second half of the 19th century, the social sciences started to arrive on the scene. During this time, people were shifting back to the ontological view of **realism**, which assumes that the physical world is 'real'; the world we perceive is *external* and *exists independently from our thought*.

The *epistemological* view was becoming more *'positivistic'*, meaning that scientists thought that we can gain knowledge about the true nature of the world through observation and experimentation. This realistic, positivistic view was mostly applied to natural phenomena. But as the social sciences developed and became distinct scientific fields, the question rose whether the realistic view should also be applied to social and psychological phenomena.

According to the view called **objectivism**, the ontological position of realism *does* indeed apply. Psychological and social phenomena like 'intelligence' and 'social cohesion' are *external, independent* properties that exist separately from our mental representation of these properties.

Objectivism can be contrasted with **constructivism**. According to constructivism, the nature of social phenomena depends on the social actors involved. This means reality is *not* independent and external; instead, reality is considered primarily a *mental construction* that depends on the observer and the context.

For example, properties like 'happiness' or 'femininity' are not external, not unchanging and cannot be objectively defined. How these properties are perceived and what they mean depends on what culture and social group the observer is part of, and the specific historical period.

So if our psychological and social reality is constructed, subjective and elusive, how do we obtain any knowledge about it? What epistemological position fits the ontological position of constructivism? Well, in fact there's a group of related views, called **Interpretivism**.

These **interpretivist** views all assume that a researcher's experience or observation of a social phenomenon can be very different from how the people who are involved in the social phenomenon experience it themselves. The focus should therefore lie with understanding the phenomenon from the point of view of the people involved.
The three interpretivist views I want to discuss are called **hermeneutics, phenomenology and verstehen**. They differ slightly on how this understanding of psychological and social reality can be gained.

Let's look at **hermeneutics** first. The term hermeneutics comes from the theological discipline concerned with the interpretation of scripture. Hermeneutics aims to explain social phenomena by *interpreting* people's behavior within their social context. Researchers need to take context into account and try to understand how people see the world in order to understand their actions.

Phenomenology is closely related to hermeneutics. It starts from the premise that people are not inert objects. They think and feel about the world around them, and this influences their actions. To understand their actions it is necessary to investigate the meaning that they attach to the phenomena that they experience.
This means investigating how people experience the world from their perspective. And to achieve such an understanding of someone else's experiences, researchers need to eliminate as many of their own preconceived notions as they possibly can.

Verstehen is the third interpretivist view. It has close ties with Hermeneutics and Phenomenology. Verstehen is mainly associated with sociologist Max Weber. Verstehen refers to the empathic understanding of social phenomena. Researchers need to assume the perspective of the research subjects to interpret how they see the world. Only then can a researcher try to explain their actions.
For example, if European researchers investigate 'happiness' in an isolated Amazonian tribe, they should do so from the tribe's perspective, taking the tribe's social context into account. For this tribe, it might be that the community is more important that the individual. This could mean that happiness is considered a group property that does not even apply to individuals. Now in order to grasp such a totally different view of the world, researchers need to immerse themselves in the culture of the person or group they are investigating.

Now of course there are some problems with the constructivist, interpretivist view. First, there is the problem of **layered interpretation**. The

researcher interprets the subject's interpretations, and then interprets the findings again as they're placed in a framework or related to a theory. With every added layer of interpretation there is more chance of misinterpretation.
A second, more serious problem is the **lack of comparability of outcomes**. When in our example happiness is subjective and means different things in different cultures we just cannot compare them. This means we can never come up with general theories or universal explanations that apply to more than just particular groups in particular periods in time.
A third problem is a difference in **frame of reference.** If the frame of reference of the researcher is very different, it can be hard for the researcher to assume the subject's point of view. This makes it hard to find out what the relevant aspects of the social context even are.

The **constructivist-interpretivist** view is generally associated with a **qualitative** approach to science. That means observations are made through unstructured interviews or **participatory observation**, where the researcher becomes part of a group to observe it.
The data are obtained from one or just a few research subjects. The data are analyzed qualitatively by interpreting texts or recorded material.

In contrast, the **objectivist – positivist** view is associated with **quantitative** research methods. Observations are collected that can be counted or measured, so that data can be aggregated over many research subjects. The subjects are intended to represent a much larger group, possibly in support of a universal explanation. The data are analyzed using quantitative, statistical techniques.

Now although a *qualitative* approach is usually associated with a *constructivist* view of science and a *quantitative* approach with an *objectivist* view, there is no reason to limit ourselves to only qualitative or only quantitative methods.
Both approaches have their advantages and drawbacks. For some research questions a qualitative approach is better, in other cases a quantitative approach is more appropriate. In fact, a **mixed-method approach**, where both methods are used to complement each other, is steadily gaining popularity.

1.10 Origins: Goals

Of course the ultimate, general goal of science is to gain knowledge. But we can distinguish more specific **goals**. These goals differ in terms of the *type of knowledge* we want to obtain and for what *purpose* we want to obtain it.

Universalistic research tries to provide explanations that apply generally. For example, we could hypothesize that playing violent computer games leads to aggressive behavior. The specific game, or the type of person playing it, is not relevant here, because we assume the relation between violent game play and

aggression holds for any violent game, be it GTA, Call of Duty, any other game. We also assume the relation holds for men and women, of any age, in any cultural setting. Universalistic research aims to describe or explain phenomena that apply to all people or all groups, or societies.

The scientific method can also be used for **Particularistic** research purposes. Particularistic research is aimed at describing or explaining a phenomenon that occurs in a specific setting or concerns a specific group.
For example, we could investigate the change in the number of Dutch teenagers hospitalized for alcohol poisoning just after the legal drinking age was first raised from 16 to 18 years in the Netherlands.
The point here is to investigate the size of an effect for a specific group in a specific location, during a very specific time. We wouldn't necessarily expect to find the same effect in a different country or in ten years time, if the drinking age was changed again.

Ok, so the goal of research can either be universalistic or particularistic. Or in less fancy terms: aimed at obtaining general versus specific knowledge. A very closely related and largely overlapping distinction is between **fundamental** and **applied** research.
Applied research is directly aimed at solving a problem. It develops or applies knowledge in order to improve "the human condition".
Suppose we want to help depressed people and we think that depression is caused by loneliness. We could create a program that aims to lower depression by making people less lonely. We could give lonely depressed people a cat to take care of and investigate if their feelings of depression actually go down now that they're no longer lonely.

Applied research can be contrasted with **fundamental research**. In fundamental research, the aim is to obtain knowledge just "for the sake of knowing"; the only purpose of fundamental research is to further our understanding of the world around us, nothing more. It doesn't have an immediate application; it doesn't directly solve a problem.
For example, we might investigate the relation between loneliness and depression in a large-scale survey study, to see whether people who feel lonelier also feel more depressed, and vice versa.

The aim here is to show there is a *relation* between loneliness and depression. Maybe we want to see if this relation exists for both men and women and for different cultural and age groups. But note that we do *not* state how depression can be treated. The goal is to know more about the relationship, not to help depressed people.

Most *fundamental* research is *universalistic*. But in some cases fundamental research *can be particularistic*, for example when research is done in a very specific setting. For example, we could investigate the relation between playing violent computer games and aggressive behavior in a very specific group of young delinquent first offenders in Amsterdam who all come from privileged backgrounds.
This very specific problem group could provide interesting insight into the relation between violent game play and aggression. Note that we're not

investigating how to rehabilitate or prevent recidivism in this group.

Applied research is often *particularistic*, aimed at solving a problem for a specific group, in a specific context, but it *can be universalistic*. Take the cat-intervention aimed at lowering depression. We could expand this applied research study by comparing a group of people that take care of a friendly cat that seeks their company and a cat that avoids any contact.
This helps us to find out more specifically what type of treatment is effective. But it also adds a universalistic element: we can investigate what it means to be lonely. Is the mere presence of a living being enough, or is interaction required?

In many cases applied research produces results that lead to new insights. These insights can be related to the intervention or treatment, but they can also provide 'fundamental knowledge'. So the two types of research can reinforce each other.

Module 2: Scientific method

2.01 Scientific method: Empirical cycle

The **empirical cycle** captures the process of coming up with hypotheses about how stuff works and testing these hypotheses against empirical data in a *systematic* and *rigorous* way. It characterizes the *hypothetico-deductive* approach to science. Being familiar with the five different phases of this cycle will really help you to keep the big picture in mind, especially when we get into the specifics of things like experimental design and sampling.

So we'll start with the **observation phase**, obviously. This where the magic happens. It's where an observation, again obviously, sparks the idea for a new research hypothesis. We might observe an intriguing pattern, an unexpected event; anything that we find interesting and that we want to explain. How we make the observation really doesn't matter. It can be a personal observation, an experience that someone else shares with us, even an imaginary observation that takes place entirely in your head.
Of course observations generally come from previous research findings, which are systematically obtained, but in principle anything goes.
Ok, so let's take as an example a personal observation of mine: I have a horrible mother-in-law. I've talked to some friends and they also complain about their mother-in-law. So this looks like an interesting pattern to me. A pattern between type of person and likeability! *Ok, so the observation phase is about observing a relation in one or more specific instances.*

In the **induction phase** this relation, observed in specific instances, is turned into a general rule. That's what induction means: taking a statement that is true in specific cases and inferring that the statement is true in all cases, always. For example, from the observation that my friends and I have horrible mothers-in-law I can induce the general rule that all mothers-in-law are horrible.
Of course this rule, or **hypothesis**, is not necessarily true, it could be wrong. That's what the rest of the cycle is about: Testing our hypothesis. *In the induction phase inductive reasoning is used to transform specific observations into a general rule or hypothesis.*

In the **deduction phase** we deduce that the relation specified in the general rule should also hold in new, specific instances. From our hypothesis we deduce an explicit expectation or prediction about new observations. For example, if all mothers-in-law are indeed horrible, then if I ask ten of my colleagues to rate their mother-in-law as either 'likable', 'neutral' or 'horrible', then they should all choose the category 'horrible'.
Now in order to make such a prediction, we need to determine the research setup. We need to decide on a definition of the relevant concepts, measurement instruments, procedures, the sample that we'll collect new data from, etcetera, etcetera.

So in the deduction phase the hypothesis is transformed by deductive reasoning and specification of the research setup into a prediction about new empirical observations.

In the **testing phase** the hypothesis is actually tested by collecting new data and comparing them to the prediction. Now this almost always requires statistical processing, using descriptive statistics to summarize the observations for us, and inferential statistics to help us decide if the prediction was correct.
In our simple example we don't need statistics. Let's say that eight out of ten colleagues rate their mother-in-law as horrible but two rate her as neutral. We can see right away that our prediction didn't come true, it was refuted! All ten mothers-in-law should have been rated as horrible. *So in the testing phase new empirical data is collected and - with the aid of statistics - the prediction is confirmed or disconfirmed.*

In the **evaluation phase** we interpret the results in terms of our hypothesis. If the prediction was confirmed this only provides *provisional support* for a hypothesis. It doesn't mean we've definitively proven the hypothesis. Because it's always possible that in the future we will find somebody who just loves their mother-in-law.
In our example the prediction was actually refuted. This doesn't mean we should reject our hypothesis outright. In many cases there are plausible explanations for our failure to confirm.
Now if these explanations have to do with the research setup, the hypothesis is preserved and investigated again, but with a better research design. In other cases the hypothesis is adjusted, based on the results. The hypothesis is rejected and discarded only in very rare cases. *In the evaluation phase the results are interpreted in terms of the hypothesis, which is provisionally supported, adjusted or rejected.*

The observations collected in the testing phase can serve as new, specific observations in the observation phase. This is why the process is described as a cycle. New empirical data obtained in the testing phase give rise to new insights that lead to a new run-through. And that's what empirical science comes down to: We try to hone in on the best hypotheses and build our understanding of the world as we go through the cycle again and again.

2.02 Scientific method: (Dis)confirmation

We're going to take a look at how we should interpret results that confirm or disconfirm our predictions and whether we should confirm or reject our hypothesis accordingly. Let's consider the hypothesis that all mothers-in-law are horrible. I formulated this hypothesis based on personal observations.

To test the hypothesis I came up with a research setup. I selected to measure horribleness using a rating scale with the options likable, neutral and horrible. I also decided to collect data from ten colleagues in my department. With the research setup in place, I formulated the following prediction: If the hypothesis

'all mothers-in-law are horrible' is true then all ten colleagues should choose the category 'horrible' to describe their mother-in-law.

Ok, let's look at **confirmation** first. Suppose all ten colleagues rated their mother-in-law as horrible. The prediction is *confirmed*, but this doesn't mean the hypothesis has been proven! It's easily conceivable that we will be proven wrong in the future. If we were to repeat the study we might find a person that simply adores their mother-in-law.
The point is that confirmation is never conclusive. The only thing we can say is that our hypothesis is **provisionally supported**. The more support from different studies, the *more credence* we afford a hypothesis, but we can never prove it. Let me repeat that: *No scientific empirical statement can ever be proven once and for all*. The best we can do is produce overwhelming support for a hypothesis.

Ok, now let's turn to **disconfirmation**. Suppose only eight out of ten colleagues rate their mother-in-law as horrible and two actually rate her as neutral. Obviously in this case our prediction turned out to be *false*. Logically speaking, empirical findings that *contradict* the hypothesis should lead to its *rejection*. If our hypothesis states that all swans are white and we then find black swans in Australia, we can very conclusively reject our hypothesis.

In practice however, especially in the social sciences, there are often *plausible alternative explanations* for our failure to confirm. These are in fact so easy to find that we *rarely reject* the hypothesis outright. In many cases these explanations have to do with methodological issues. The research design or the measurement instrument wasn't appropriate, maybe relevant background variables weren't controlled for; etcetera, etcetera.

Coming back to our mother-in-law example: I could have made a procedural error while collecting responses from the two colleagues who rated their mother-in-law as neutral. Maybe I forgot to tell them their responses were confidential, making them uncomfortable to choose the most negative category.
If there are plausible methodological explanations for the failure to confirm we *preserve* the hypothesis and instead choose to reject the *auxiliary, implicit assumptions* concerning the research design and the measurement. We investigate the original hypothesis again, only with a better research setup.

Sometimes results do give rise to a **modification** of the hypothesis. Suppose that the eight colleagues who did have horrible mothers-in-law were all women and the other two were men. Perhaps all mothers-in-law are indeed horrible, but only to their daughters-in-law!

If we alter the hypothesis slightly by adding additional clauses (it only applies to daughters-in-law), then strictly speaking we are rejecting the original hypothesis, sort of. Of course the new hypothesis is essentially the same as the original one, just not as general and therefore not as strong. An outright rejection or radical adjustment of a hypothesis is actually very rare in the social sciences. Progress is made in very small increments not giant leaps.

2.03 Scientific method: Criteria

We follow the empirical cycle to come up with hypotheses and to test and evaluate them against observations. But once the results are in, a confirmation doesn't mean a hypothesis has been proven and a disconfirmation doesn't automatically mean we reject it.

So how *do* we decide whether we find a study convincing? Well, there are two main criteria for evaluation: **reliability** and **validity**.
Reliability is very closely related to *replicability*. A study is replicable if independent researchers are in principle able to repeat it. A research finding is *reliable* if we actually repeat the study and then find consistent results.
Validity is more complicated. A study is *valid* if the conclusion about the hypothesized relation between properties *accurately reflects reality*. In short, a study is valid if the conclusion based on the results is 'true'.

Suppose I hypothesize that loneliness causes feelings of depression. I deduce that if I decrease loneliness in elderly people, by giving them a cat to take care of, their feelings of depression should also decrease. Now suppose I perform this study in a retirement home and find that depression actually decreases after residents take care of a cat. Is this study valid? Do the results support the conclusion that loneliness causes depression?
Well, because this is still a pretty general question, we'll consider three more specific types of validity: **construct**, **internal** and **external** validity.

Construct validity is an important prerequisite for internal and external validity. A study has high construct validity if the properties or constructs that appear in the hypothesis are measured and manipulated accurately. In other words, our methods have high construct validity if they actually *measure* and *manipulate* the properties that we *intended* them to.
Suppose I accidently measured an entirely different construct with for example my depression questionnaire. What if it measures feelings of social exclusion instead of depression? Or suppose taking care of the cat didn't affect loneliness at all, but instead increased feelings of responsibility and self-worth. What if loneliness remained the same?

Well, then the results only *seem* to support the hypothesis that *loneliness* caused depression, when in reality we've manipulated a different cause and measured a different effect. Developing accurate measurement and manipulation methods is one of the biggest challenges in the social and behavioral sciences. I'll discuss this in more detail, when we look at operationalization.

For now I'll move on to **internal validity.** Internal validity is relevant when our hypothesis describes a *causal relationship*. A study is internally valid if the observed effect is *actually due* to the hypothesized cause.

Let's assume our measurement and manipulation methods are valid for a second. Can we conclude depression went down because the elderly felt less lonely? Well... maybe something else caused the decrease in depression. For example, if the study started in the winter and ended in spring, then maybe the change in season lowered depression. Or maybe it wasn't the cat's company but the increased physical exercise from cleaning the litter box and feeding bowl?
Alternative explanations like these, threaten internal validity. If there's a plausible alternative explanation, internal validity is low. Now, there are many different types of threats to internal validity that I will discuss in much more detail in later videos.

OK, let's look at **external validity**. A study is externally valid if the hypothesized relationship, supported by our findings, also *holds in other settings and other groups*. In other words, if the results *generalize* to different people, groups, environments and times.
Let's return to our example. Will taking care of a cat decrease depression in teenagers and middle-aged people too? Will the effect be the same for men and women? What about people from different cultures? Will a dog be as effective as a cat?
Of course this is hard to say based on the results of only elderly people and cats. If we had included younger people, people from different cultural backgrounds and used other animals, we might have been more confident about this study's external validity. I'll come back to external validity, and how it can be threatened, when we come to the subject of sampling.

So to summarize: **Construct validity** relates to whether our *methods actually reflect* the properties we *intended* to *manipulate* and *measure*. **Internal validity** relates to whether our *hypothesized cause* is the *actual cause* for the observed effect. Internal validity is threatened by alternative explanations. **External validity** or *generalizability* relates to whether the hypothesized relation *holds in other settings*.

2.04 Scientific method: Causality

The most interesting hypotheses are the ones that describe a **causal relationship**. If we know what causes an effect we can *predict*, *influence* it, better *understand* it.

So how do we identify a causal relationship? Well, it was David Hume, with a little help from John Stuart Mill, who first listed the criteria that we still use today. These are the four essential criteria:

- Number one: The cause and effect are **connected**. The first criterion means there has to be a way to trace the effect back to the cause. If a patient was not exposed to a virus, then we can't argue that the virus *caused* the patient's death.
- Number two: The cause **precedes** the effect. I hope this is obvious.

- Number three: The cause and effect **occur together consistently**. This means cause and effect should go together, *covary*. When the cause is present, we should see the effect, and if the cause is not present, then the effect should be absent. If the cause influences the effect to a certain degree, then we should see a consistently stronger or weaker effect, accordingly.
- Criterion number four: **Alternative explanations** can be ruled out.

Ok, so let me illustrate these criteria with an example. Suppose I hypothesize that loneliness causes feelings of depression. I give some lonely, depressed people a cat to take care of. Now they're no longer lonely. If my hypothesis is correct, then we would expect this to lower their depression.
The cause and effect, loneliness and depression, *are in close proximity*, they happen in the same persons, and fairly close together in time, so we can show they're connected. The cause, a decrease in loneliness, needs to happen *before* the effect, a decrease in depression. We can show this because we can control the presence of the cause, loneliness.
The cause and effect should occur together consistently. This means that less loneliness should go together with lower depression. I could find a comparison group of lonely, depressed people that do not get a cat. Since the cause is absent: their loneliness doesn't change, there should be no effect.

This was all easy, the real difficulty lies in the last criterion, excluding any alternative explanations, other possible causes. Let's look for an alternative explanation in our example. Maybe the increased physical activity, required to take care of a cat, actually caused lower depression, instead of the reduction in loneliness. **Alternative explanations** form *threats to the internal validity* of a research study. An important part of methodology is developing and choosing research designs that minimize these threats.

Ok, there is one more point I want to make about causation. *Causation requires correlation*; the cause and effect have to occur consistently. But **correlation doesn't imply causation**!
I'll give you an example: If we consistently observe aggressive behavior after children play a violent videogame, this doesn't mean the game *caused* the aggressive behavior. It could be that aggressive children seek out more aggressive stimuli, reversing the causal direction. Or maybe children whose parents allow them to play violent games aren't supervised as closely. Maybe they are just as aggressive as other children, they just feel less inhibited to show this aggressive behavior. So remember: correlation does not imply causation!

2.05 Scientific method: Internal validity threats - participants

Internal validity is threatened if there's a plausible alternative explanation for a study's results. In order to judge the internal validity of a particular study you have to be familiar with the type of threats that can occur. I'll start with three types of threats that are in some way associated with the participants or the subjects used in the study. These threats are called **maturation**, **selection** and **selection by maturation**.

Let's start with maturation. Maturation refers to an alternative explanation formed by *natural change*. Suppose I hypothesize that loneliness causes depression. I decrease loneliness in elderly people, who are prone to depression, by giving them a cat to take care of. I expect their depression to go down because they are now less lonely.

I find a retirement home willing to participate. And I measure depression in a group of residents who seem unhappy. I give them a cat to take care of for four months and then I measure depression again. Let's assume depression is lowered after taking care of the cat. Does this mean that the cat's companionship *caused* the decrease in depression? Well, not necessarily, the decrease in depression could have occurred naturally. People develop, they change. Many physical and mental problems simply disappear after a while. Even if we don't receive treatment, depressions often go away by themselves.

Fortunately there is way to eliminate this alternative explanation of natural change that we refer to as maturation. We can introduce a **control group** that is measured at the same times but is *not* exposed to the hypothesized cause. Both groups should 'mature' or change to the same degree. Any difference between the groups can now be attributed to the hypothesized cause and not natural change.

The threat of maturation is eliminated, but unfortunately a study that includes a control group is still vulnerable to other threats to internal validity. This brings us to the threat of **selection**. Selection refers to any systematic difference in subject characteristics between groups, other than the manipulated cause.

Suppose in my study I included a control group that didn't take care of a cat. What if assignment of elderly participants to the groups was based on mobility? Suppose people who weren't able to bend over and clean the litter box were put in the control group. Now suppose the experimental group was in fact less depressed than the control group. This might be caused, not by the company of the cat, but because the people in the experimental group were just more physically fit.
A solution to this threat is to use a method of assignment to groups that ensures that a systematic difference on subject characteristics is highly unlikely. This method is called **randomization**. I'll discuss it in much more detail when we cover research designs.

The last threat to internal validity related to participants, is the *combined* threat of maturation and selection. We call this a **selection by maturation** threat. This happens when groups systematically differ in their rate of maturation. For example, suppose the effectiveness of the cat treatment was examined in an

experimental group consisting of volunteers who are open to new things. In contrast, the control group consisted of more conservative people who don't like change.

Participants were selected so that both groups had a similar level of depression at the start of the study. But what if we find that the experimental group shows lower depression? Well perhaps the lower rate of depression in the experimental, cat-therapy group is simply due to the fact that open-minded people tend to naturally "get over" their depressive feeling more quickly than conservative people do. Just like selection on its own, the threat of selection by maturation can be eliminated by *randomized assignment* to groups.

So you see that the research design we choose - for example adding a control group - and the methods we use - for example random assignment - can help to minimize threats to internal validity.

2.06 Scientific method: Internal validity threats - instruments

Another category of threats to internal validity is associated with the instruments that are used to measure and manipulate the constructs in our hypothesis. The threats of **low construct validity**, **instrumentation** and **testing** fall into this category.

I'll start with **low construct validity**. Construct validity is low if our instruments contain a systemic bias or measure another construct or property entirely. In this case there's not much point in further considering the internal validity of a study. As I discussed in an earlier video, construct validity is a *prerequisite* for internal validity. If our measurement instruments or manipulation methods are of poor quality, then we can't infer anything about the relation between the hypothesized constructs.

Suppose I hypothesize that loneliness causes depression. I attempt to lower loneliness of elderly people in a retirement home, by giving them a cat to take care of, expecting their depression to go down. Suppose that taking care of the cat didn't affect loneliness at all, but instead gave residents a higher social status: they're special because they're allowed to have a cat. The manipulation, aimed at lowering loneliness, in fact changed a different property, social status.
Now consider my measurement of depression. What if the questionnaire that I used, actually measured feeling socially accepted, instead of feeling depressed? If we accidently manipulated social status or measured 'feeling accepted', then we cannot conclude anything about the relation between loneliness and depression.

The second threat that relates to measurement methods is **instrumentation**. The threat of instrumentation occurs when an instrument is changed during the course of the study. Suppose I use a self-report questionnaire to measure depression at the start of the study, but I switch to a different questionnaire or maybe to an open interview at the end of the study.

Well, then any difference in depression scores might be explained by the use of different instruments. For example, the scores on the post-test, at the end of the study, could be lower because the new instrument measures slightly different aspects of depression, for example.

Of course it seems rather stupid to change your methods or instruments halfway, but sometimes a researcher has to depend on others for data, for example when using tests that are constructed by national testing agencies or polling agencies. A good example is the use of the standardized diagnostic tool called the DSM. This 'Diagnostic and Statistical Manual' is used to classify things like mental illness, depression, autism and is updated every ten to fifteen years.
Now you can imagine the problems this can cause, for example for a researcher who is doing a long-term study on schizophrenia. In the recently updated DSM, several subtypes of schizophrenia are no longer recognized. Now if we see a decline in schizophrenia in the coming years, is this a 'real' effect or is it due to the change in the measurement system?

The last threat I want to discuss here is **testing**, also referred to as sensitization. Administering a test or measurement procedure can affect people's behavior. A testing threat occurs if this sensitizing effect of measuring provides an alternative explanation for our results. For example, taking the depression pre-test, at the start of the study, might alert people to their feelings of depression. This might cause them to be more proactive about improving their emotional state, for example by being more social. Of course this threat can be eliminated by introducing a control group; both groups will be similarly affected by the testing effect. Their depression scores will both go down, but hopefully more so in the 'cat-companionship' group.

Adding a control group is not always enough though. In some cases there's a risk that the pre-test sensitizes people in the control group differently than people in the experimental group.
For example, in our cat-study, the pre-test in combination with getting a cat could alert people in the experimental 'cat'-group to the purpose of the study. They might report lower depression on the post-test, not because they're less depressed, but to ensure the study seems successful, so they can keep the cat!
Suppose people in the control group don't figure out the purpose of the study, because they don't get a cat. They're not sensitized and not motivated to change their scores on the post-test. This difference in sensitization provides an alternative explanation. One solution is to *add* an experimental and control group that *weren't* given a pre-test. I'll discuss this solution in more detail when we consider different experimental designs.

Ok, so to summarize, *internal validity* can be threatened by **low construct validity**, **instrumentation** and **testing**. Low construct validity and instrumentation can be eliminated by using *valid instruments* and *valid manipulation methods* and of course by using them *consistently*. Testing can be eliminated by using a *special design* that includes groups that are exposed to a pre-test and groups that aren't.

2.07 Scientific method: Internal validity threats - artifacts

Another category of threats to internal validity is associated with the 'artificial' or unnatural reaction caused by participating in a research study. This can be a reaction from the participant, but also from the researcher!
In any scientific study the researcher has to observe the behavior of individuals or groups. And in most cases, the people under observation are aware of being observed. Both researcher and participants have expectations about the goal of the study and the desired behavior. These expectations can lead to a change in behavior that can provide an alternative explanation. I'll discuss two of these threats to internal validity right here: **experimenter expectancy** and **demand characteristics**.

If the researcher's expectations have a biasing effect, we call this threat to internal validity an **experimenter expectancy** effect. Experimenter expectancy refers to an unconscious change in a researcher's behavior, caused by expectations about the outcome, that influences a participant's responses. Of course this becomes an even bigger problem if a researcher unconsciously treats people in the control group differently from people in the experimental group.

One of the most subtle and shocking demonstrations of an experimenter expectancy effect was undoubtedly provided by Rosenthal in the nineteen sixties. Psychology students were led to believe they were taking part in a course on practical research skills in handling animals and "duplicating experimental results".
Students handled one of two batches of rats that were compared on their performance in navigating a maze. The first batch of rats, students were told, was bred for their excellent spatial ability. The other batch was bred for the exact opposite purpose.

Lo and behold, there was a difference in performance; the maze-bright rats outperformed the maze-dull rats. Of course in reality there was no difference in maze-brightness in the two groups of rats, they were both from the same breeding line and they were randomly assigned to be 'bright' or 'dull'.
Apparently, just knowing what was expected of their batch of rats led students to treat them differently, resulting in an actual difference in performance of the rats. Imagine what the effect can be when a researcher interacts with human participants!

Fortunately there is a solution. If the experimenter who interacts with the participants doesn't know what behavior is expected of them, the experimenter will not be able to unconsciously influence participants. We call this an **experimenter-blind** design.

Of course the participant can also have expectations about the purpose of the study. Participants who are aware that they are subjects in a study will look for cues to figure out what the study is about. **Demand characteristics** refers to a change in participant behavior due to their expectations about the study. **Demand characteristics** are especially problematic if people respond to cues differently in the control and in the experimental group.

A well-known form of demand characteristic occurs when people are aware that they are in an experimental group and are undergoing a treatment, especially if the treatment aims to help them. Participants might be grateful to be in this group, or hopeful that the treatment will be effective. And this might lead them to be more positive about the treatment then had they been unaware of it.

But the cues don't even have to be accurately interpreted by the participants. As long as participants in the same group interpret the cues in the same manner and change their behavior in the same way, demand characteristics form a real problem.
This is why it's always a good idea to leave participants unaware of the actual purpose of the study or at least leave them unaware of which group they are in, the experimental or the control group. If both the subject and the experimenter are unaware, we call this a **double-blind** research design.

Because *any* cue can lead to a bias in behavior, researchers usually come up with a **cover story**. A cover story is a plausible explanation of the purpose of the study. It should provide participants with cues that are unlikely to bias their behavior. Of course this temporary deception needs to be *necessary*, the risk of bias due to demand characteristics needs to be real. And of course you need to debrief participants afterwards and inform them about the real purpose of the study.

2.08 Scientific method: Internal validity threats - design/procedure

The last category of threats to internal validity is a bit of a remainder category. Very generally the three types of threats in this category are related to the research procedure or set-up. The threats are **ambiguous temporal precedence, history** and **mortality**.

An **ambiguous temporal precedence** in the hypothesized causal relation is just a fancy way of saying that it is unclear if the hypothesized cause actually *precedes* the observed effect. Suppose I'm interested in the relationship between playing violent videogames and aggressive behavior. I ask high school students how many hours a week they play violent games and I ask their teacher to rate their aggressiveness in class.
What if I find a strong relation: Children who play violent games for many hours a week also show more aggressive behavior? Well, this *doesn't* mean violent game play causes aggressive behavior. Maybe children who play more violent games were more aggressive to begin with and are more likely to seek out violent stimuli.

The threat of ambiguous temporal precedence can be eliminated by manipulating or introducing the hypothesized cause. Of course not all constructs can be manipulated.
But if I can manipulate the cause I can make sure it happens *before* the effect. For example, I can make *all* children play violent games. If children that were not aggressive to begin with also become more aggressive *after* playing the violent game, then my causal inference is much stronger.

Let's move on to a threat referred to as **history**. A history effect is an unforeseen event that happens during the study that provides an alternative explanation. This could be a large-scale event or something small that goes wrong during data collection.
Consider a study on mitigating negative stereotypes about a minority group. The manipulation consists of a group discussion, led by an experimenter. The experimenter focuses on the point of view of the minority group, asking participants to put themselves in their shoes. In the control condition the experimenter focuses on differences between the majority and minority and stresses the point of view of the majority. In both groups there are three weekly group discussions.

Ok, to give an example of a history effect on a small scale, imagine that during the last session in the control group, the experimenter faints. Of course participants are shaken and upset about this. And this might translate into a more general negative attitude in the control group, which also makes the control group's attitude towards the minority more negative. The treatment might look effective, because the experimental group is more positive, but the difference is due, not the discussion technique, but due to the fainting incident.

Let's consider the history effect on a larger scale. Suppose that during the study a horrific murder is committed, allegedly by a member of the minority. The crime gets an enormous amount of media attention, reinforcing the negative stereotype about the minority group. Any positive effect of the intervention could be undone by this event.

The threat of history is hard to eliminate. Large-scale events, well they can't be avoided. Small-scale events that happen during the study can be avoided, at least to some extent, by testing subjects separately, if this is possible. This way, if something goes wrong, the results of only one or maybe a few subjects will have to be discarded.

The final threat to discuss is **mortality**. Mortality refers to participant *dropout* from the study. If groups are compared and dropout is different in these groups, then this could provide an alternative explanation. For example, suppose we're investigating the effectiveness of a drug for depression. Suppose the drug is actually not very effective, and has a very strong side effect. It causes extreme flatulence!
Of course this can be so uncomfortable and embarrassing that it wouldn't be strange for people to drop out of the study because of this side effect. Suppose 80% of people in the experimental group dropped out. In the control group participants are administered a placebo, with no active ingredient, so also, no side effects. Dropout in this group is only 10%. It's obvious that the groups are no longer comparable.

Suppose that for the remaining 20% of participants in the experimental group the drug is effective enough to outweigh the negative side effect. This wasn't the case for the 80% who dropped out. Based on the remaining subjects we might conclude that the drug is very effective. But if all subjects had remained in the study the conclusion would have been very different.

The threat of mortality is very hard to eliminate, in most cases the best a researcher can do is document the reasons for dropout so that these reasons can be investigated and possibly mitigated in further studies.

2.09 Scientific method: Variables of interest

I want to go into research designs and give you a more concrete idea how a research design can minimize threats to internal validity. But before I can do that you have to become familiar with the terms **construct**, **variable**, **constant** and **independent** and **dependent variable**.

A hypothesis describes or explains a relationship between *constructs*. The term **construct** is used to indicate that we are talking about a property in general, abstract terms. For example, I could hypothesize that loneliness and depression are associated. The terms loneliness and depression are the *constructs* here.

Of course loneliness and depression can be expressed in many different ways. The term **variable** refers to an **operationalized version of a construct**. A variable is a specific, concrete expression of the construct and is **measurable** or **manipulable**. For example, I could operationalize loneliness in a group of elderly people in a nursing home for example by using a self-report questionnaire. I could administer the UCLA Loneliness Scale. This scale is a 20-item questionnaire consisting of items like: ″ I have nobody to talk to″. The variable loneliness now refers to loneliness as expressed through scores on the UCLA-scale.

If I hypothesize that loneliness *causes* depression, I would be better off manipulating instead of measuring loneliness. I could give one group of elderly people a cat to take care of, comparing them with a control group without a cat. I have now operationalized loneliness by creating two levels of loneliness. The variable loneliness now refers to 'high' or 'low' loneliness expressed through the presence or absence of a feline companion.
Finally, I could operationalize depression by using the Geriatric Depression Scale, the GDS, consisting of 15 questions such as “Do you feel happy most of the time?”. The variable depression now refers to depression as expressed through scores on the GDS.

Ok, so variables are measured or manipulated properties that take on different values. This last bit is important, a variable's values need to vary, otherwise the property isn’t very interesting. Suppose the nursing home is so horrible that all residents get the maximum depression score. Well then we cannot show a relation between loneliness and depression, at least not in this group of subjects. Both lonely and less lonely people will be equally depressed. Depression is a **constant**.
So the *variables* central to our hypothesis should be variable, they should show variation. Of course it *is* good idea to keep o*ther, extraneous* variables constant, so that

they cannot provide alternative explanations, but we'll get to that in another video.

Ok now that I've defined what a variable is, let's look at different types of variables according to the role they play in describing or explaining a phenomenon. I'll refer to the variables that are central to our hypothesis as **variables of interest**.

When a cause and effect can't be identified or when a causal direction isn't obvious or even of interest, our variables are 'on equal footing'. Then we just refer to them as variables. But when our hypothesis is causal, we can identify a cause and an effect. And we then refer to the cause variable as the **independent variable** and to the effect variable as the **dependent variable**.

The *independent* variable is also referred to as **cause** variable, **explanatory** or **input** variable or **predictor**. It refers to a variable that is hypothesized to cause or predict another variable. In our example loneliness was hypothesized to cause depression. The independent variable here of course is loneliness, and it's operationalized through the presence or absence of a cat.

The *dependent* variable is hypothesized to be influenced by the cause variable or to be the result of another variable. Its values *depend* on another variable. In our example the *dependent* variable was depression, as measured through scores on GDS questionnaire. The dependent variable is also referred to as **effect** variable, **response** variable, **outcome** or **output** variable.

Now if you're having trouble telling the terms independent and dependent apart, try to remember that the independent variable is what the researcher would like to be **in** control of, it's the cause that comes f**ir**st.

2.10 Scientific method: Variables of disinterest

I've discussed variables that are the focus of our hypothesis: the *variables of interest*. Of course in any study there will be *other, extraneous* properties associated with the participants and research setting that vary between participants.
These properties are not the main focus of our study but they might be associated with our variables of interest, providing possible alternative explanations. Such **variables of disinterest** come in three flavors: **confounders**, **control variables** and **background variables**.

A **confounder** or lurking variable, is a variable that is related to both the independent and dependent variable and partially or even entirely accounts for the relation between these two. Suppose I investigate the effect of reducing loneliness on depression in a group of elderly people. I lower loneliness by providing a cat to the elderly in an experimental group; and elderly in a control group receive a stuffed toy.

Now, besides loneliness, the two groups might also differ in terms of the physical exercise they get, their age or their susceptibility to depression, which are all *variables of disinterest*. Take physical exercise. The experimental group will likely be more physically active because they have to feed and clean up after the cat. So physical exercise is related to loneliness, the cat group – the less lonely group - is more active than the control group.

Suppose physical activity is also causally related to depression - being more active lowers depression. Well, then physical activity, and not loneliness, may account for a lower depression score in the experimental cat group. The relation between loneliness and depression is said to be **spurious**. The relation can be explained by the *confounding variable*, physical activity.

An important thing to note about confounders is that they are not included in the hypothesis and they're generally not measured. This makes it impossible to determine what the actual effect of a confounder was. The only thing to do is to repeat the study and control the confounder by making sure it takes on the same value for all participants. For example, if all elderly people in both groups are required to be equally active, then physical activity cannot explain differences in depression.

Another possibility is to turn a confounder into a control variable. A **control variable** is a property that is likely to be related to the independent and dependent variable, just like a confounder. But unlike a confounder, a control variable *is* measured. Its effects can therefore be assessed and controlled for.
For example, we could see if physical activity provides an alternative explanation by measuring it and taking it into account. Suppose we can distinguish inactive and active people. In the cat-therapy there are more active people, but some are inactive. In the control condition most people are inactive, but some are active.

We now consider the difference in depression between the cat-therapy and control group, first for the active people, and then for the inactive people. We control for activity, in fact holding it constant by *considering each activity level separately*.
If the relationship between loneliness and depression disappears when we look at each activity level separately, then activity 'explains away' the *spurious* relation between loneliness and depression. But if the relationship still shows at each activity level, then we have eliminated physical activity as an alternative explanation for the drop in depression.

The last type of variable of disinterest is a **background** variable. This type of variable is not immediately relevant to the relation between the variables of interest, but it is relevant to determine how representative the participants in our study are for a larger group, maybe all elderly everywhere, even all people, of any age. For this reason it is interesting to know how many men and women participated, what their mean age was, their ethnic or cultural background, social economic status, education level, and whatever else is relevant to the study at hand.

So to summarize: a **confounder** is a variable that partially or entirely *explains* an effect on the dependent variable instead of, or additional to the independent variable. A confounder is *not accounted for* in the hypothesis and is *not measured or controlled* for in the study. A possible confounder can be controlled for by keeping the property *constant* or by turning it into a control variable.
A **control variable** accounts for a possible confounder by *measuring* the relevant property and checking the relationship between the variables of interest, at each value or level of the control variable.
Background variables finally, are measured, not because a possible effect on the variables of interest is expected, but because the background properties are useful to assess the *generalizability* of the study based on the sample characteristics.

Module 3: Research designs

3.01 Research designs: True experiments

Hypotheses that claim a causal relationship are very interesting, but also very bold and, especially in the social sciences, very susceptible to alternative explanations or threats to internal validity.
Experimental research designs maximize internal validty. They are referred to as **true experiments**, also known as **randomized control trials** or **RCT's**. They are science's best defense against alternative explanations for causal claims. The three *essential* ingredients of a true experiment are **manipulation, comparison** and **random assignment**.

Let's start with **manipulation**. If you want to show a causal relation, the strongest possible empirical demonstration is one where the cause is under *your* control.
If you can *create* a situation where the cause is present, a causal relation is more plausible, because you can show that it *precedes* the effect, eliminating an ambiguous temporal precedence.

What about **comparison**? Well causality is even more plausible if you can *compare* to a situation where the cause is absent, showing that the effect does not occur when the cause is absent. This also eliminates the threat of maturation.

Think of the relation between violent imagery and aggression. Let's say I measure how many hours a week a group of ten year olds play violent video games and how aggressive they are, according to their teacher. Suppose I find a positive relationship: Kids who play more violent videogames tend to be more aggressive. I can argue that playing games increases aggression. But of course, I can also argue that aggressive children seek out more violent stimuli.

I could have approached this problem differently and encouraged a subgroup of children to play a violent videogame (say GTA V) for a total of ten hours in one week and deny the other group any access to violent imagery for the same period of time. Now if I find that the children who've been denied access to violent imagery are less aggressive than the group who played the violent game regularly, then I have a stronger case for the causal claim that violent imagery causes aggression.

Of course it's not a *very strong* case, since there are still many alternative explanations for the difference in aggression between these two groups. What if the kids in the video game group were more aggressive to start out with? What if there were fewer girls in this group or more older children?

This is where **randomization** comes in. I can randomly assign children to the experimental condition with heavy video-play or the control condition with no access to violent imagery, and I can do this by flipping a coin: Heads

for the experimental condition, tails for the control condition. *On average*, this process will ensure an equal distribution over the two groups in terms of gender, but also in terms of age, previous aggression, hair color, shoe size, I can go on.

On average, randomization ensures that there is no systematic difference between the groups other than the difference in the independent variable, the cause variable. Of course in any one particular study it is possible, entirely due to chance, that randomization fails and we end up, for example, with more girls in the control group, possibly explaining why this group is less aggressive. The only way to be sure randomization worked is to replicate a study and show that the results are consistent each time!

So to summarize: *manipulation* ensures the cause precedes the effect, *comparison* to a control group ensures the effect did not occur naturally and *random assignment* ensures that there are no other systematic differences between the groups that could explain the effect. *Replication* is generally not considered a characteristic of a true experiment, but it is required to ensure randomization actually works.

3.02 Research designs: Factorial designs

An important type of *experimental* research design is the **factorial design**. In a factorial design several *independent* variables, also called **factors**, are investigated simultaneously.

For example, we could investigate the effectiveness of an experimental drug aiming to reduce migraine attacks. Suppose we create three conditions that differ in the administered dosage: low, medium and high. We can now investigate the effect of the factor dosage on the number of migraine attacks: is a higher dosage more effective in reducing attacks?
We can extend this simple design and make it factorial by adding a second factor, for example gender. If we make sure there are enough, preferably equal numbers of men and women assigned to each of the dosages, then we end up with six conditions: Men who receive a low, medium or high dosage, and women who receive a low, medium or high dosage.

Besides the effect of dosage, we can now also investigate the effect of the second factor gender: Do women suffer from more migraine attacks than men? We can also investigate the combined effect of dosage and gender. We can see whether a higher dosage is more effective in reducing the number of migraine attacks for women as compared to men.
The effects of the factors *separately* are called **main effects**. The *combined* effect is called the **interaction effect**. In this case we're dealing with a *two-way interaction*, because the effect combines two factors.

Of course we can add more factors, making it possible to investigate *higher-order interactions*. But as the number of factors increases the design becomes very complicated very quickly. Suppose we add diet as a factor, with two conditions: a normal diet and a diet eliminating all chocolate and red wine, let's call this the no-fun diet. This would require that each of the six groups is split in two, with half of the participants being assigned to the normal diet and half to the no-fun-diet.
We can now look at the *main effect* of diet: is a no-fun diet effective at reducing migraine attacks?

We can also look at the *two-way interaction* between diet and gender, maybe the no-fun diet is effective for men but not for women? We can also look at the *two-way interaction* between diet and dosage: Is a higher dosage more effective when a no-fun diet is being followed as compared to a normal diet?

Finally, we can look at the *three-way interaction*. Maybe a higher dosage is effective for women regardless of diet, but maybe for men a higher dosage is effective only if they follow a no-fun-diet and not if they follow a normal diet. Like I said, it can get pretty complicated pretty quickly.

There are even more complicated factorial designs, called *incomplete designs* where not all combinations of levels of the factors, or **cells**, are actually present in the design. Now, we won't go into those designs right now.

What's important here is that you understand the basics of factorial designs. You need to know that a factorial design consists of two or more independent variables and - for now - one dependent variable. The independent variables are *crossed*, to ensure that all *cells* are represented in the study and that each cell contains enough participants. Factorial designs are very useful because they allow us to investigate not only the main effects of several factors, but also the combined effects, or interaction effects.

3.03 Research designs: Repeated measures

When we investigate an independent variable that can be manipulated, the standard approach is to expose different groups of participants to the different levels of the independent variable. We call this a **'between-subjects design'** and the independent variable a between subjects factor, or simply, a **between factor**.

In some cases it's possible to let each participant experience all levels of the independent variable. When participants are exposed to all conditions, we refer to this as a **'within subjects' design**. The independent variable is now called a ***within* subjects factor** or just a **within factor**.

Suppose we investigate the effectiveness of an experimental drug in reducing migraine attacks at different dosages. Our participants are migraine

patients. A standard approach would be to assign patients randomly to receive a low, medium or high dosage of the drug for one week. But we could also choose to let each participant experience all three dosages, one after the other, for a total of three weeks.

Within factors can be combined with other within factors or between factors in a factorial design. For example, in addition to the within-factor dosage, we could also investigate the factor gender. Of course the participants can't be exposed to both levels of the variable gender, so gender remains a between factor. But it can be combined with the within factor dosage: A group of men are exposed to all three dosages, and so are a group of women. This allows us to investigate whether different dosages of the drug are more effective for women than for men.

Now we could have investigated the independent variables dosages and gender using a between-between design, with different men and women each in different dosage conditions. But a within factor is more efficient for statistical, but also for practical reasons.

If you need forty participants in each condition, it might be easier to find just forty people who are willing to participate for three weeks than it is to find one hundred and twenty people willing to participate for one week.

The concept of a within factor is closely related to the terms **repeated measure design** and **longitudinal design**. Both terms refer, obviously, to studies where the dependent variable is measured repeatedly.
In a within subjects design, the same participants are measured on the dependent variable after being exposed to each level of the independent variable. Otherwise we wouldn't know what the effect of each level or condition is.
The term **'repeated measures design'** is used as a synonym for a within subjects design, but is also used to refer to more complex designs with at least one within factor and possibly one or more between factors.

The term **'longitudinal design**' refers to studies that measure the same variables repeatedly, over a long period of time. We're talking months, even years or decades.
The term 'longitudinal design' usually refers to correlational studies where no independent variables are manipulated. But the term does include experimental or quasi-experimental studies that succeed in long-term manipulation of independent variables; such studies are rare though.

So a longitudinal study repeatedly measures variables over a long period. Repeated measure designs also measure repeatedly, but run shorter and refer to studies with manipulated independent variables where there is at least one within factor and possibly one or more between factors. The term within-subjects design is generally used to indicate a design consisting of within factors only.

3.04 Research designs: Manipulation

Manipulation is one of the essential ingredients of a true experiment. Manipulation generally refers to control over the independent variable. In a true experiment the value or level of the independent variable that a participant experiences, is determined - or **manipulated** - by the researcher.
It also helps to control external variables. By keeping variables of disinterest constant we can rule out any alternative explanations they might have otherwise provided.

Let's start with manipulation of the independent variable. Suppose we hypothesize that violent imagery is a direct cause of aggression. In order to test this hypothesis we could manipulate the independent variable 'violent imagery', by letting participants play a violent video game for either two hours, four hours or not at all.
In this case we've created three **levels** of the independent variable 'violent imagery'. The term 'levels' nicely indicates that the independent variable is present in different ways or to different degrees in these three settings.

Other frequently used terms are **'conditions'** and **'groups'**. If the independent variable is absent this level is called the **control condition** or **control group**. In our example this is the group that does not play the violent video game.

If the independent variable is fully controlled by the researcher it is often referred to as an **experimental variable**. Not all variables can be manipulated. Such variables are called '**individual difference variables**', because they are an intrinsic property of the participant. Properties like age or gender, for example, are not under the researchers' control. We can't send participants to a gender clinic and ask them to undergo a sex change, so that we can investigate the effect of gender on aggression.

However, some variables that *seem* like non-manipulable, individual difference variables *can* be manipulated. For example, a variable like self-esteem could be manipulated by giving participants a bogus intelligence test. In one condition participants are told they scored extremely high, thereby boosting self-esteem.
In another condition participants are told they scored far below average, decreasing their self-esteem. We can now investigate the effect of high or low self esteem on a subsequent math test for example.

It is important to realize that manipulation can fail. Maybe the test was very easy and therefore participants in the experimental condition didn't believe their scores were low, leaving their self-esteem unaffected! We can check whether the intended level of the independent variable was actually experienced, by measuring it. This is referred to as a **manipulation check**. It is important to perform this check, this measurement *after* measuring the dependent, effect variable. Otherwise you might give away the purpose of the experiment. Asking participants about their self-esteem before the math test might lead them to question the feedback they've received.

Let's move on to control of the variables of disinterest. In the ideal case, each condition is entirely identical to the others except for the independent variable. This is referred to as the "Ceteris Paribus" principle. It means "all other things equal".
Suppose all other properties are the same, or constant, and only the independent variable differs between conditions. If we find an effect - a difference between conditions on the dependent variable - then we can assume this effect is caused by the independent variable.

Now in physics it is relatively easy to keep "All other things equal". We can control external variables - like temperature or air friction - to make sure these properties don't change between individual observations and between conditions. A social scientist's job is *much* harder. It's impossible to control all socially and psychologically relevant aspects of a situation. Not only are there a lot more properties that could provide alternative explanations; it is often much harder to keep these variables under control.

Variables that are held constant are called **control variables**. In our video game study we could make sure that the wall color is the same in all conditions. We wouldn't want the wall to be a calming blue in the control condition and a bright, agitating red in the two violent gaming conditions, for example. It becomes much harder when the variables of disinterest are individual differences variables like a participant's age or a country's average educational level. This is where randomization and matching come in. But I'll discuss both of these later.

So to summarize: Manipulation is about creating different levels or conditions that represent different values of the independent variable. Effectiveness of a manipulation can be assessed using a manipulation check. Control means keeping external variables the same across conditions. For individual differences variables manipulation or experimental control is not possible.

3.05 Research designs: Lab vs. field

A rigorous investigation of a causal hypothesis requires manipulation of the independent variable and control of extraneous variables, keeping them all the same across conditions.

This type of *experimental study* requires a large amount of control. Control over the setting and circumstances under which the research is conducted. This is why a lot of experimental research is done in a **laboratory** or **lab**. The experimental method combined with the control that a lab setting offers, maximizes internal validity.

Now in the social sciences a lab isn't a room full of test tubes and microscopes. It's simply an environment that's entirely under the researchers control. A lab room could be a small room with no distracting features, just a comfortable chair and a desk with a computer. So that participants can perform, for example, a

computerized intelligence test without distraction and all under the same conditions. Or it could be a room with soft carpeting, colorful pillows and toys, fitted with cameras to record, for example, the reaction of small children when a stranger enters the room.

The lab is very useful for experimental studies, but that doesn't mean that lab studies are by definition always experimental, they can also focus on non-causal hypotheses without any manipulation, just using the lab to control extraneous variables.

Ok, so lab research generally has high internal validity, but some argue that it has low **ecological validity**. Ecological validity, or *mundane realism*, refers to how closely the lab setting approximates how people would naturally experience a phenomenon.

Suppose we want to investigate the effect of low self-confidence on negotiating skills. Participants in our study are given extremely complicated instructions and asked if they understand. In all cases the instructions are repeated more clearly, but in the experimental group this remark is made first: "You seem confused, you don't understand these instructions? Wow…". Where the 'wow' subtly implies that the participant isn't the brightest bulb in the box. Obviously, this remark isn't made in the control group.

The participants then take part in a computer simulated salary negotiation. They are asked to image they are going to start a new job. And they get the first salary offer displayed on the screen. They are asked to select a counter-offer from one of four options, or agree to the offered salary. If they are too aggressive in their negotiation the offered salary will go down.

Now of course this setup doesn't approximate a real salary negotiation with a face to face meeting, where non-verbal behavior can play a role and substantive arguments are used in the negotiating process, obviously.
But low ecological validity like this isn't necessarily bad. It doesn't automatically imply low *construct* and *external* validity. In a lab setting researchers try to find and experimental translation of the phenomenon as it occurs naturally. This is referred to as **experimental realism**.

Simulating the negotiating process using the computer with very limited choices and no face-to-face contact is highly artificial. But within the lab setting this procedure might suffice to demonstrate that lower self-confidence is in fact related to accepting a lower salary offer. Similarly, in real life most people wouldn't become less self-confident just based on one subtle derogatory statement about their intelligence made by a stranger. But in the lab setting, with the experimenter in a position of power and the participant likely to feel judged and vulnerable, this manipulation might actually be very appropriate and a very effective way to induce short-term lower self-confidence.

The experimental translation might be very different from what happens in 'real life', but that doesn't mean that within the lab setting, the construct isn't adequately manipulated or measured and that the lab results won't generalize to other, more 'natural' instances of the investigated relationship.

Of course research can also be done in the **field**, meaning outside the lab in an uncontrolled environment, like a public area or a private residence. Field research naturally lends itself to the observation of natural behavior 'in the wild', but field research can be experimental.

For example, we could repeat the study on the effect of self-confidence on negotiating success in the field with a group of candidates selected for a traineeship program at a large bank. All candidates undergo one final assessment with lots of very difficult tests. We could then tell one half of the group that they scored below average on the assessment and the other half that they scored above average and we can just see how well they do in their salary negotiations. Of course such a study would be highly unethical. And there would be all kinds of variables we wouldn't be able to control for. But both types of studies have their advantages and disadvantages and they can complement each other, one type maximizing internal validity and the other maximizing external validity.

3.06 Research designs: Randomization

Randomization, or **random assignment**, provides a way of eliminating all possible systematic differences between conditions all at once. Think of the relation between violent imagery and aggression. Suppose I encourage a group of children to play a violent videogame (say GTA V) for a total of ten hours in one week and I deny another group any access to violent imagery for the same period of time.

Suppose I find that children who played the violent game are more aggressive than the control group. Of course there are still many possible alternative explanations for the difference in aggressive behavior other than violent stimuli. Suppose children could volunteer for the violent game-play condition. It would not be hard to image that this group would consist of more boys, or children more drawn to violence and aggression than the control group.

Such systematic differences between the groups - providing alternative explanations for more aggressiveness in the experimental group - would likely have been prevented with random assignment. Ok, let's see why.

I could have randomly assign children to the conditions by flipping a coin: heads for the experimental condition, tails for the control condition. A naturally aggressive child would have a fifty-fifty chance of ending up in the experimental condition. The same goes for a child of the male sex, a very meek child, a child with impaired eyesight, a child with large feet. Any property you can think of, a child with that particular property will have an equal chance of being assigned to one of the conditions.

Put another way: How many boys do we expect in the experimental condition? About half of all boys in the study, because for the boys - on average - the coin will show heads half of the time. How many naturally aggressive children will end up in the experimental condition? Again, about the same number as in the control condition. And the same goes for all other characteristics we can think of.

On average, randomization ensures that there is no systematic difference between the groups other than on the independent variable. Of course in any one *particular* study it is possible - entirely due to chance - that we end up, for example, with more girls in the control group, possibly explaining why this group is less aggressive.

I call this a **randomization failure**. We rely on the law of large numbers when we flip a coin, but this law doesn't always work out, especially in small groups. Suppose there are only 4 boys and 4 girls to assign to the two groups. It's not hard to imagine that the coin toss will come out something like this:

[first girl] - heads - [second girl] - tails - [first boy] - heads - [second boy] - heads - [third boy] - heads - [third girl] - heads - [fourth girl] - tails - [fourth boy]- tails.

The experimental group now consists of 3 boys and 2 girls, five children in all. The control group consists of 1 boy and 2 girls, three children in all. The problem is that the groups are not of equal size, which is a nuisance statistically speaking. We also have a systematic difference in terms of sex.

One solution is to perform a **randomization check**. We can measure relevant background or control variables and simply check to see whether randomization worked and the conditions are the same, or whether randomization failed and the conditions differ on these variables.

There is a way to guarantee randomization works on a select set of variables, by using **restricted randomization** procedures. Blocking is the simplest form of restricted randomization. It ensures equal or almost equal group sizes. We pair children up in blocks of two and flip a coin to determine where the first child goes. The second child is automatically assigned to the other condition. Repeat for all children and - if we have an equal number of participants - equal group sizes are ensured.

Now in **stratified restricted random assignment** we use the blocks to ensure not just to equal numbers, but also equal representation of a specific subject characteristic, for example equal numbers of boys and girls in each group. We can arrange this by first pairing up all the girls and for each block of girls flipping a coin to determine to what condition the first girl is assigned. We then automatically assign the second girl to the other condition.

We now have a girl in each condition; we do the same for the second block of girls and end up with two girls in each condition. The same method is applied in assigning the boys so that we end up with two boys and two girls in each condition.

Of course stratified randomization has its limits, you can apply it to several characteristics combined, sex and age for example, but with more than two or three variables to stratify on, things become complicated. Moreover, there's an endless number of subject characteristics, it's impossible to control them all.

I have one final remark about randomization in repeated measures designs, concerning within-subjects designs, where all subjects are exposed to all of the conditions. In this case randomization might seem unnecessary, but this is not the case!

If all subjects are exposed to the conditions in the same order, then any effect could be explained by maturation, or some sort of habituation effect that spills over from one condition to the other.

So in within-subjects designs, the *order* in which subjects are exposed to the conditions should be randomized. We call this **counterbalancing**. Subjects are assigned to one out of all possible orderings of the conditions, possibly using blocking to ensure that no one ordering is overrepresented.

3.07 Research designs: Experimental designs

A true experiment is the best way to maximize internal validity. The key elements of a true experiment are *manipulation* of the independent variable, *comparison* between conditions exposed to different levels of the independent variable and of course *random assignment* to these conditions.

Of course these elements can be implemented in very different ways. I'll discuss four experimental designs that are very common. The simplest design is the **two-group design**. Participants are randomly assigned to one of two conditions, usually an experimental condition where the hypothesized cause is present and a control condition where it's absent.

The independent variable could also differ between the conditions in amount or kind, for example if we're investigating the effect of male versus female math teachers on math performance of boys, for example. In the two-group design the dependent variable is measured after exposure to the independent variable to assess the difference between conditions, which are likely to be similar in all respects due to the random assignment, including their pre-existing position on the dependent variable.

Of course in small groups, randomization doesn't always work. In such cases it might be wise to use a **two-group pretest/posttest design**, which adds a pretest of the dependent variable before exposure to the independent variable. With a pretest you can check whether for example both groups of boys were equally proficient at math before being exposed to a female versus a male math teacher for a month. This is an

especially good idea when maturation forms a plausible threat to internal validity.

A pretest also allows the researcher to compare the size of the increase or decrease in scores in the experimental and control condition. For example, we can assess how much the boys' math performance increased due to natural improvement and what the additional effect of teacher sex was.
Unfortunately a pretest can sometimes sensitize participants. The pretest may result in a practice effect, leading to higher scores on the posttest or it may alert participants to the purpose of the study. Especially if this effect is stronger for one of the conditions internal validity will be negatively affected. But there is a way to take such unwanted effects of a pretest into account by using a **Solomon four-group design**. This is a combination of the two-group design and the two-group pretest/posttest design. The experimental and control condition are run twice, once with a pretest and once without.

For example, it is possible that the math test isn't very hard to begin with and provides good practice in those math skills that the boys still lack. On the posttest the boys in both conditions get perfect scores, obscuring any effect that teacher sex might have. If we had two other groups of boys that didn't take the pretest we might see the effect of teacher sex, because these groups have had less practice.

Of course if we find a difference between these groups it could still be attributable to an already existing difference in math proficiency, but together with the results of the pretest groups we could come up with a better, more difficult test, showing differences between the two pretest groups and two non-pretest groups in a follow-up study.

Another very common design is the **repeated measures design** with one **within-subjects factor**. In this design all participants are exposed to all levels of the independent variable, they experience all conditions.

For example we could randomly select half of the boys to have a female math teacher for a month and then a male teacher the following month. The other half of the boys would be taught by the male teacher during the first month and the female teacher during the second month.

The only thing that is really different to the previous, between-subjects designs is that the random assignment of participants is not to the conditions themselves, because they experience all of them, but to the order in which the conditions are experienced.

3.08 Research designs: Matching

In many situations it's not possible to assign participants to conditions randomly. Of course the threat of selection to internal validity automatically applies in that case. Systematic differences between conditions are now much harder to rule out.

Random assignment can be impossible due to pragmatic or ethical reasons, but also when the independent variable is an *individual differences variable*. For example, if we want to investigate the effect of sex on political conservativeness, we can't randomly assign people to be female or male.

When random assignment is impossible, one way to mitigate the selection threat to internal validity is to **match** participants on relevant background variables. We find **matching** groups on these variables and discard participants that do not match up. For example, we could match men and women in terms of age and maybe educational level, to make sure that they don't differ systematically, at least on these two properties. We have thereby excluded two possible alternative explanations for any difference in political conservativeness between men and women.

There is a potential danger in the use of matching though! A thing called **undermatching** could occur. This can happen when the conditions are matched on a variable that is measured with some degree of error and is related to the variables of interest. To understand what this means and how undermatching can occur you first need to understand what **regression to the mean** is.

Suppose we were able to measure someone's intelligence repeatedly without any practice effect. Assuming the test we use is valid and reliable, we would still get slightly different result each time, since we cannot measure intelligence perfectly. Ok, now suppose we pick a random person from our pool of participants. We measure their intelligence and find a very high score, of 132. The mean intelligence score is 100, and about 70% of people score between 85 and 115. So what kind of score would you expect if we measured this person's intelligence again?

Well because such a high score is very uncommon, it's more likely that the score of 132 is an overestimation of someone's 'real' intelligence. The next score for this person will probably be lower. Of course we could have a genius in our pool of participants and find a higher score on the second test. But if we look at a group of people with high scores, we can say that on average their second score will be lower. Some will get a higher score, but most will get a lower score, closer to the mean. So that's why we call this change in scores regression toward the mean.

How can this cause problems when we use matching to create comparable conditions? Well suppose we want to investigate the effectiveness of watching Sesame Street in improving cognitive skills for disadvantaged versus advantaged toddlers.

Let's say our disadvantaged toddlers come from poor, broken homes, they receive very little stimulation to develop cognitive skills like reading and counting. The advantaged children have nurturing parents and are provided with ample educational stimulation.

If we want to investigate if watching Sesame Street improves cognitive skills differently for disadvantaged versus advantaged children, then it would seem like a good idea to match children at the start of the study in terms of cognitive ability, using a pretest. If we start with groups of equal ability we can get a good idea of the effect of Sesame Street for both groups. If one group starts out smarter, then the improvement in both groups is just harder to compare.

Now, matching is only a problem if the variable that we match on is related to our variables of interest. Here we use a pretest of cognitive ability to select comparable disadvantaged and advantaged toddlers.
So in this case the relation between the matching variable and dependent variable is very strong, because they measure the same property. It's also highly likely that our matching variable - cognitive ability, the pretest - is related to the independent variable, advantaged versus disadvantaged background. It is likely that 'in real life' children who lack stimulation and security - whether through nature or nurture - already have lower cognitive abilities.

What happens if these groups differ in cognitive ability and we select a sample of toddlers from the disadvantaged and the advantaged group so that they have about the same cognitive ability? We match them up. Well that means we choose disadvantages toddlers with relatively high scores, relative to the mean score of their group and advantaged children with relatively low scores, relative to the mean score of their group.

Now in the disadvantaged selection it's likely that at least some of these relatively high scores are overestimations and a second measurement will be closer to the mean of the disadvantaged children, resulting in a lower mean score for this group. In the advantaged selection it's likely that at least some of these relatively low scores are underestimations and a second measurement will be closer to the mean of the advantaged children, resulting in a higher mean score for this selection.

So without any intervention we would already expect a difference between these groups on the post-test, based on just the regression to the mean effect. Of course this effect might lead to horribly inaccurate conclusions about the effectiveness of watching Sesame Street.

Suppose the effect of watching Sesame Street was small but equally effective for both groups. A large regression effect could result in lower scores for disadvantaged kids, leading us to conclude that watching Sesame Street is bad for disadvantaged children. This distorting effect of regression to the mean due to matching, showing a *detrimental* effect instead of the hypothesized beneficial effect is called **undermatching**.

Now this effect can only occur if the matching variable is related to the variables of interest *and* is measured with a fair amount of error - which is unfortunately the case for most social and psychological variables. This of course does not apply to variables like age, sex and educational level, which can be assessed without, well almost without, measurement error.

3.09 Research designs: Quasi-experimental designs

Quasi-experimental designs look and feel like experimental designs but they always lack the key feature of random assignment. They can lack the feature of manipulation and comparison as well. Like a true experiment, a quasi-experiment is generally used to demonstrate a causal relationship. In some cases the researcher can manipulate the independent variable but is unable to assign people randomly due to practical or ethical reasons.

For example, suppose we want to investigate the effect of playing violent computer games on aggressive behavior of young children. We might not get permission from the parents, unless they can decide for themselves whether their child is exposed to a violent game or a puzzle game. This parent-selected assignment is obviously not random.

In other cases the independent variable is an individual differences variable and can't be manipulated. Or it's a variable that happens 'naturally' to some people and not others, like a traumatic war experience. Some people refer to studies that employ these types of independent variables as 'natural experiments'.

Suppose that we also wanted to consider differences in aggressiveness between boys and girls. This automatically entails a quasi-experimental design, since we cannot assign children to be a boy or a girl for the purpose of our study. Studies where the independent variable wasn't manipulated but 'selected' are in fact very similar to correlational studies, which I'll discuss later, the difference is that quasi-experimental studies examine causal hypotheses and exert as much control as possible over extraneous variables.

Ok so let's look at some quasi-experimental designs. When the independent variable can be manipulated or 'selected', conditions can be compared. The simplest design at our disposal in that case is the **static group comparison design**. This design is similar to the two-group experimental design with just a posttest after exposure to the independent variable.

The only difference is non-random assignment, which means the selection threat to internal validity is always present. The comparison of groups lessens the threats of maturation and history. But it is very well possible that a pre-existing difference between conditions on the dependent variable exists. This

selection threat also makes the causal direction more ambiguous.

What if more permissive parents, with more aggressive children selected the violent game, showing a higher aggressiveness to begin with? Did violent stimuli cause aggression or did permissiveness and aggression lead to the selection of violence? Of course this problem could be fixed by extending the design to a **pretest/posttest nonequivalent control group design** simply by adding a pretest.
With this design we're able to determine whether there were any pre-existing differences on the dependent variable. If we can show that such differences don't exist, we've firmly disambiguated the causal order. We can also assess the size of any maturation effects.
Unfortunately it's not always possible to use a control group. In that case the only thing you can do is at least establish temporal precedence of the cause by using a one-group pretest/posttest design. All the other threats to internal validity still apply though.

One way to improve on the one group pretest/posttest design is to include more measurements of the dependent variable before and after exposure to the independent variable. We refer to this as an **interrupted time-series design**. More observations before and after the 'treatment' allow you to assess any natural change in the dependent variable. If the change is gradual with a sudden jump from just before to just after the 'treatment', we can rule out - at least to some extent - the threats of maturation and history.

Suppose we only have access to the violent-game group but we have measurements at several time points for them. Aggressiveness increases slightly during the entire study. The change is the same just before and after exposure to the violent game, indicating that there was just no effect on aggressiveness. Now consider these results: Aggressiveness doesn't change until the violent game is played. Immediately after, aggressiveness remains stable at a higher level, indicating a long-lasting effect.

Things could also look like this: Aggressiveness increases slightly until after the violent game is played. And the results show a jump in aggressiveness only after exposure and the same slight increase after that. This might indicate a natural increase and effect of violent game-play. Of course it would be wise to check whether any other event occurred right at the sudden jump that might form a history threat to the internal validity.

One way to check whether there truly is an effect and not a history effect is to use a **replicated interrupted time-series design**. This design adds a second group of participants for whom the dependent variable is measured at the same time points, but no 'treatment' is administered. This is basically a more complicated version of the pretest/posttest nonequivalent control group design.

Consider this outcome: if we see the same pattern for the second group there's probably a history threat present. The design could also be implemented by exposing a second group to the same 'treatment' but at a different time than the first group. Consider this outcome: if the effect also shows if the intervention

is presented just a little bit later, we can be more sure it was actually caused by the violent game-play.

3.10 Research designs: Correlational designs

The term **correlational design** refers either to studies that do not employ any form of manipulation of the independent variable, or to studies that don't even identify an independent variable, because the hypothesis doesn't specify a causal relationship.
So in correlational studies we don't manipulate or select, we just measure.
Now if an independent variable is identified, any causal inference needs to be made with extreme caution, because temporal precedence of the cause is really hard to establish, never mind all the other threats to internal validity, especially selection.

Just like with experimental and quasi-experimental designs, there are a couple of standard correlational designs that you should be familiar with. First of all, there's the **cross-sectional design** in which a cross-section of a population, one - usually large – group, is considered at one specific point in time. Usually a fairly large number of properties are measured at once.

The aim is to investigate the association between variables in a way that accurately describes a larger population. Within the large sample, groups may be compared but it's important to note that participants were not selected beforehand to represent a level of an independent variable like in quasi-experimental studies with individual differences variables.
The term survey study or survey design is sometimes used to denote a cross-sectional design, since these studies often make use of surveys. But this is an unfortunate term, because the method of measurement really has nothing to do with the research setup.

Another type of study is a **time-series design.** A time-series design can refer to one person being measured at several points in time, usually with many measurement moments in quick succession and in fixed intervals. In some social science fields the term time-series design is used interchangeably with the term longitudinal design. But 'longitudinal design' generally refers to a group being measured at several points in time, so this can lead to confusion.
The term time-series is also used for quasi-experimental designs where one or more conditions are measured repeatedly before and after an intervention or placebo is administered.

The term **panel design** or time-series cross-sectional design, is used for non-experimental studies that follow more than one individual over a longer period of time, without an intervention or placebo. In panel designs the same group of people is measured at several points in time. A special case of a panel design is a cohort design, where the participants are all the same age, or started a new school or job at the same time.

Ok, it's easy to get these terms mixed up with the terms longitudinal and repeated measures designs. Generally speaking longitudinal refers to any study that follows one or more participants over a *long* period of time, whether it's experimental, quasi-experimental or correlational.
Time-series design usually refers to correlational studies of just one person measured at fixed intervals. The term time-series can also refer to groups of individuals being measured repeatedly, but then the term is associated with quasi-experimental designs.
The term time-series design is sometimes also used for experimental studies, although the term N=1 study is more popular in this context. The term repeated measures implies an experimental study where at least one independent variable was manipulated or selected.

So to summarize: In correlational designs researchers distinguish three types of studies that differ on the dimensions of individuals and time. Cross-sectional designs concern the measurement of many individuals - usually on many variables - at one point in time. Time-series-designs follow only one individual over several points in time. Panel studies combine both dimensions, by considering a group of the same individuals at several points in time.

3.11 Research designs: Other designs

You have to be familiar with the standard experimental, quasi-experimental and correlational designs. But you will also encounter some special types of studies that are less common or used only in specific fields.

Let's start with case studies. The term **case study** refers to studies that focus on one person or one group. You're already familiar with case studies at least the ones that are quantitatively oriented and have an experimental, quasi-experimental or correlational design. They're referred to as single-subject, time-series or N=1 research. The term case study is more often associated with qualitative studies, which are generally aimed at generating hypotheses instead of testing them.

Now I won't go into purely qualitative case studies here, because this course focuses on the hypothetico-deductive approach to science. But there is a type of qualitative case study that actually fits this approach perfectly. It's called **negative case analysis**.
Negative case analysis means that the researcher actively searches for a case that provides contradictory evidence, evidence against a hypothesis. Now supporting a hypothesis requires a lot of consistent confirmatory evidence and it's always provisional. But in theory, you just need one good counter-example to reject a hypothesis. Now of course social and psychological hypotheses usually specify relationships that apply in general or on average, or to groups, not individuals. So occasional negative cases normally don't lead to rejection. But if a hypothesis is

formulated in all or none terms negative case analysis can be very useful.

Another type of study is **evaluation research**, aimed at investigating the effectiveness of a policy or program implemented to address a specific problem. This type of research can be **summative**, focusing on the outcome of a program, assessing whether it was effective, or it can be **formative**, focusing on the process, assessing how the program works.
Evaluation research is a form of applied research, since it investigates a program implemented 'in the real world' and it's not necessarily, but usually non-experimental, because often it's just impossible to use a control group, due to ethical or practical reasons.

Intervention studies on the other hand are usually experimental. They're aimed at investigating the effectiveness of a method aimed at treating problems of individuals, generally in a clinical setting.
Think of studies on cognitive behavioral therapy for depression or remedial teaching for children with dyslexia. In contrast, evaluation research focuses on programs with a broader scope, aimed at larger societal or educational problems.
Evaluation research is common in sociology, communication and political sciences; intervention studies are more the domain of developmental and clinical psychologists.

Validation research is another very specific type of research. This research is aimed at assessing the quality of a measurement instrument. The instruments are usually surveys, questionnaires or tests designed to measure an attitude, a trait or an ability. The instruments typically consist of several questions that are supposed to all tap into the same property of interest. Statistical analysis is used to see if the responses to these questions show high internal consistency.

Another major topic of analysis in validation studies, is whether responses to questions that measure related but slightly different properties show expected patterns of association. Validation studies are important in the field of psychometrics and sociometrics.

Module 4: Measurement

4.01 Measurement: Operationalization

Before we turn to the topic of measurement I'll briefly clarify the terms variable and operationalization. Earlier I referred to variables as operationalized constructs, but the term **variable** can also refer to a representation of a construct that is still somewhat abstract. We use the term **operationalization** if we want to explicitly indicate we're talking about a specific, concrete method to measure or manipulate a construct.

Say I'm interested in the construct 'political power'. I can represent this construct with the variable 'influence in parliament'. This representation is more specific, but we could still come up with very different procedures to operationalize or measure the variable 'influence in parliament'.

For example, I could count the number of bills that a Member of Parliament, or Congress, got passed, or I could count the number of years someone has been in parliament, or I could ask political analysts to rate members of parliament in terms of their influence.

So operationalization means selection or creation of a specific procedure to measure or manipulate the construct of interest. An operationalization makes it possible to assign people an actual score on the variable of interest.

Suppose I want to operationalize the construct 'love of animals'. I can observe a person interacting with a cat and count how often the person pets or strokes the cat. I could also decide on a different operationalization or **operational definition** by creating a questionnaire with statements like 'I usually avoid other people's pets' and 'I love animals'.

What about independent variables that are manipulated? Well, suppose I want to know if exposure to animals increases love of animals. I can operationalize the variable 'exposure to animals' by creating two levels of exposure. I could randomly assign people to take care of a cat from a shelter for a month or assign them to a control condition with not cat.

Another operationalization would be to take one half of a school class to a petting zoo and the other half to an abstract art museum. Or I could assign participants to watch an animal documentary or a train documentary. As you can see, the possibilities for operationalization are endless. Of course some operationalizations are better than others.

An operationalization doesn't necessarily capture or represent the construct in its entirety. As our constructs get 'fuzzier' or more complex, there's a greater chance that we measure or manipulate only a part of the construct. For example, if we measure love of animals with a self-report questionnaire, we measure feelings and attitudes, which might give a more positive image. If we measure love of animals by placing camera's in people's homes and observing behavior we might find lower scores.

We might find that, compared to their self-reported love of animals, people show a lot less love when their cat wakes them up at five AM or blocks the TV when they're watching they're favorite show.
It's important to keep in mind what aspect of the construct the operationalization actually measures or manipulates, especially once we use the data to draw conclusions about the entire construct in our hypothesis. Our conclusion may apply only to a limited aspect of the construct.

4.02 Measurement: Measurement structure

Once a construct has been operationalized we're ready to start measuring. Unfortunately, measurement is much less straight forward in the social sciences than in the natural sciences. Therefore it is extremely important to know to what we mean when we say we're *measuring* depression or political persuasiveness. We should know what we're capturing with the numbers that result from measurement, but more importantly, we should know what information we're *not* capturing.

So what is measurement? Ok, here we go: Measurement is the representation of relations between objects, persons or groups on a certain property by using relations between numbers. Take the property body length. I can determine qualitative relations on this property between three 'objects' - in this example: two persons and a cat- by standing them side-by-side and observing and comparing how far their heads stick out.

The first thing I can tell is that they're all of different length. I could represent this inequality relation by using different labels or numbers to represent the inequalities. Of course it would be weird to use the numbers 2, 10 and 14 like this, because there is another type of relation that we can immediately see, which is *not* reflected in the assigned numbers.
I'm talking about the order relation between A, B and C. Person B is the tallest, so he should receive the highest number, 14, and C is shortest and so should receive the lowest number, 2. We use the ordering of the numbers to represent the ordering of people - and cat- in terms of body length.

And we don't need to stop there. We can also determine if the difference in length between person A and person B is the same or larger than the difference between person A and C. We hold a piece of cardboard over A's head an cut the cardboard where it reaches the top of B's head. Then we hold the piece of

cardboard over C's head and compare to A. Suppose the cardboard reaches exactly to A's head, then the differences in length are the same.
We can represent this relation of 'equal differences' by using numbers that differ by the same amount. For example, the difference between the numbers for B and A, 14 and 10 is four, so we could change the number assigned to C from a 2 to a 6.
The equal differences in numbers between B and A, 14-10=4, and A and C 10-6=4, now accurately reflect the equal differences in body length between A and B and A and C. If the difference between A and C had been larger than the difference between A and B, then the difference in the corresponding numbers should have reflected this.

There's one more type of relation we can observe for the property body length. We can compare ratios of body length. We could take the piece of cardboard we cut out earlier, cut some extra pieces of exactly the same length and see how many cardboard units C, A and B are tall.

Ok, now suppose it takes two cardboard units to reach the top of C's head, four to reach the top of B's head and three to reach the top of A's head. This means B is twice as tall as C.
We could reflect this relation by changing the numbers again to 9, 12 and 6. We're still using different numbers for different lengths, the ordering of the numbers corresponds to the ordering of body lengths, the differences between A and B and A and C are the same (twelve minus nine is three and nine minus six is also three) and now the number for person B is twice as large as the number for C.

So you can see that measurement is the representation of empirical relations, between objects or persons on a certain property, by using the numerical relations between numbers. We can differentiate lengths, order them and compare differences and ratios of body length. We determined these empirical relations by looking and using some cardboard. Of course this method is pretty laborious if we want to assess body length for a great number of people.

Assigning numbers in a way that captures these empirical relations, for example by using a tape measure, makes our life a lot easier; especially if we want to compare or aggregate over many people. And of course assigning numbers to represent a property allows us to use statistics to help describe and draw conclusions about the property we are interested in.

4.03 Measurement: Measurement levels

We saw that measurement is the representation of relations between people on a certain property by using corresponding relations between numbers. Consider body length. For this property we can distinguish different body lengths, we can order them, compare differences in body length and even compare ratios. We can use the numerical relations of numbers to represent all these relations.

For body length all four possible relations - inequality, order, differences and ratios - can meaningfully interpreted. Unfortunately this is not the case for psychological and social properties. For most properties in the social sciences we can only determine some of these relations. The term **measurement level** is used to indicate what type of relation can be meaningfully interpreted.

If the only relation you can determine is that of inequality, distinguishing between values, then we call this a **nominal** variable. An instrument that can only differentiate between values is said to have a **nominal measurement level**. Examples are nationality, sex or pet preference.
A German has a different nationality than a Brit, women are of a different sex than men, dog people have a different preference for pets than cat people or hamster people. One value doesn't represent a greater degree of the property than any other value, they're just different. A German doesn't have more nationality than a Brit, women don't have more sex, well, let's say more gender than men and being a dog person doesn't mean you have a stronger animal preference than a cat or hamster person. There's no order to these properties.

Ordinal variables allow for differentiation *and* ordering of values. Suppose I want to measure math ability and use the number of correct answers on a math test with ten questions. The higher someone's math ability, the more answers they get right. We can order people's test scores to reflect their order in math ability, but differences or ratios of scores don't reflect differences or ratios of math ability.

We have no way of showing that the difference between a score of four and five is the same in terms of different math ability as the difference between a score of seven and eight. Sure, the difference in right answers is the same but how can we show this corresponds to an equal difference in math ability? We can't. And the same goes for ratios: someone with a score of ten doesn't have twice the mental math ability of someone with a score of five.

It actually remains to be seen if the test scores are measured at the ordinal level, that is, if they accurately reflect the order in math ability. What if someone with a score of one spent all their time on the hardest question and got it right, where someone else focused on the easier questions and got a score of three?

Only if the questions are equally difficult, can we use the test scores to accurately

reflect the ordering of students on math ability. In that case the math test is said to measure at the **ordinal level**.

For **interval** variables it's possible to distinguish and order values, but also to interpret *differences* between values. Temperature is a good example. Suppose I'm heating up four pans filled with water on a stove and I measure temperature with a thermometer in degrees Fahrenheit. A pan of water reading 90 degrees Fahrenheit is hotter than one that reads 80; we can verify this by sticking our hand in. And the same goes for two pans reading 40 and 50 degrees.
We can also verify that when we heat up the 80 degree water to 90 degrees the expansion of a liquid, like the quicksilver in a thermometer, is the same as the expansion when we heat up water at 40 degrees to 50 degrees. So the difference between 80 and 90 and 40 and 50 is the same.

We can't say however, that water at 80 degrees Fahrenheit is twice as hot as water at 40 degrees. This is because the zero point for temperature is arbitrarily defined. The value 0 doesn't correspond to the absence of temperature, it corresponds to the temperature required to freeze brine, or salt water. The Celsius scale defines zero as the temperature at which fresh water freezes.
If we consider the same temperatures as before but now in degrees Celsius we see that thirty-two point two minus twenty-six point six is five point six, just like ten minus four point four is five point six. But twenty-six point six is nowhere near twice four point four. This is because the scales use different zero points.

Unlike interval variables, **ratio** variables have a non-arbitrary zero point that's the same for any scale you might choose. Of course length is an obvious example. The absence of length, 0 length, is the same whether you measure in inches or in centimeters. Variables measured at the interval or ratio level are very rare in the social sciences.

On final remark: The structure of a property doesn't have to be fully captured by a measurement instrument. Take age, a ratio property. I could measure age by asking respondents to indicate their age in years, thereby preserving the ratio level.
I could also ask them whether they are under twenty, twenty to thirty-nine, forty to fifty-nine or sixty or older, assigning the scores one, two, three and four. By creating age categories we no longer know exactly how old someone is. We can say that people in a higher category are older, but not by how much. By categorizing the variable we've lost the ratio and interval level information.

4.04 Measurement: Variable types

We identified the measurement levels nominal, ordinal, interval and ratio. In this video I'll look at how you can interpret variables with different measurement levels. I'll also discuss other ways to classify variables according to their measurement characteristics.

Categorical variables distinguish either unordered or ordered categories. A special type of categorical variable is a **binary**, or **dichotomous** variable. This type of variable has exactly two categories, such as male or female, smoker or non-smoker, furry or hairless. The categories can be natural, like the male/female dichotomy or created by the researcher such as under twenty years of age and twenty or older. Categorical variables with more than two categories are sometimes called **polytomous**.

For categorical variables differences between values are uninterpretable. Of course nominal and ordinal variables are both categorical. So how can we interpret numerical results from categorical variables? Well, suppose I measure animal preference in a group of my friends by assigning the numbers one, two and three respectively to friends who prefer dogs, cats, or hamsters.
It doesn't make sense to look at the mean animal preference of, say, 1.2 and say that my friends have a low preference for animals. I could have assigned the numbers in the reversed order, resulting in a high mean of 2.8. For a nominal variable like animal preference it only makes sense to look at frequencies, how many people there are in each category.

What about a math test with ten questions that measures math ability at the ordinal level? Suppose I administer the test to my friends and find a mean score of 6.2. Is this mean interpretable? Well not if the scores only reflect ordering. Because I could reassign the person with the highest score the number 15. The ordering is still the same so the relations are preserved. Of course if I did this, the average test score be suddenly much higher.
The value is arbitrary and not informative of real differences between people on the property of interest. So if you have an ordinal variable you should stick to frequencies and statistics like the mode and the median.

Categorical variables can be contrasted with **quantitative** variables. Quantitative variables allow us to determine not only that people differ on a certain property, but also to what extent they differ. Interval and ratio variables are quantitative.

For quantitative variables like temperature, weight and length it does make sense to calculate a mean and, for example, to compare means of groups. This is because the mean is influenced by the distance between numbers. For quantitative variables the distance between numbers actually corresponds to the distance between people on the property of interest.

For example, if I measure the weight of friends who own a cat and the weight of my friends who own a dog, I can compare their mean weight to see

if cat people are heavier because they don't get the extra exercise from walking their pet.

A final distinction that you should be able to make is between discrete and continuous variables. For continuous variables it's always possible, in theory anyway, to find a value between any other two values. Consider body weight: If one person weighs 65 kilograms and another 66 kilograms, we can easily image finding someone who weighs 65.5 or 65.72 or 65.268334. As long as our measurement instrument is precise enough, any value between the minimum and maximum on the scale should be possible.

Discrete variables on the other hand, can only take on a limited set of values. Nominal and ordinal variables are by their nature discrete. But quantitative variables can also be discrete. Take the number of pets someone has owned. This is a ratio variable, because differences can be compared: the difference between two and four pets is the same as between one and three pets, and because ratio's can be compared: someone with four pets owns twice as many pets as someone who with two pets.
The set of possible values is very limited however. We start at 0 and then the values, 1, 2, 3, 4. But 1.3 pets or 4.7 pets are not valid values. So here we have an example of a discrete ratio variable.

The distinction between continuous and discrete variables is less relevant, because it's not associated with how you can interpret the measurement results unlike the distinction between categorical and quantitative variables.

4.05 Measurement: Measurement validity

Until now I've discussed operationalization and measurement levels without asking "Are we measuring accurately, do our measurements reflect the construct we are interested in?" in other words: is our instrument valid? The validity of an instrument or manipulation method is commonly referred to as **measurement** or **construct validity**.

How do we assess construct validity? Well, suppose I've created a questionnaire that aims to measure fondness of cats. A higher score indicates someone is more of a cat person. How do we determine if this score actually reflects the property 'fondness of cats'?
Well, we could determine its **face validity**. Face validity refers to how well the instrument represents the property according to the assessment of experts. An expert opinion can be useful in the development phase, but of course experts can always be wrong. A slightly better alternative is to show the instrument has **predictive validity** or **criterion validity**, by demonstrating the instrument can predict a relevant property or criterion. Of course the ability to predict something doesn't mean the scores used for prediction accurately reflect the intended construct.

Suppose I create an instrument to measure motivation that can predict job satisfaction. That doesn't mean the instrument measures motivation, it could reflect another construct entirely, say general positive attitude. So **criterion validity** can support the claim that we are measuring *something* consistently, but it has limited value in demonstrating that this is indeed the intended construct.
What would be ideal is if we already had a valid instrument for the property of interest. We could then administer both instruments and see whether the scores on the new scale agreed with the already validated scale. Unfortunately there aren't many gold standard instruments for social and psychological constructs.

Another solution would be if we could directly check our measurements. Consider body length: We can use a tape measure and then check if the person whose head sticks out furthest gets the highest measurement result. This purely qualitative way to assess a property is cumbersome, but it allows us to directly check the validity of a tape measure or a bathroom scale.
For social and psychological constructs the situation is very different. We don't have an undisputed, direct way to determine whether one person is more intelligent or fonder of cats than another. So is there another way to assess construct validity? Well we can go about it indirectly, by seeing whether the scores relate to similar and different variables in a way that we expect. We refer to this as **convergent** and **discriminant validity**.

For example, I would expect scores on my cat fondness scale to show agreement, or converge, with scores on an observational measure of cat fondness, where people spend ten minutes in a room with a cat and we count the number of times the person looks at or pets the cat.
I would expect less agreement between my cat fondness questionnaire and a questionnaire on fondness of wild animals. It wouldn't be strange to find some association, but I would expect it to be smaller. Finally, I would expect no association with a variable that is supposedly unrelated to cat fondness, like fondness of pizza.

A systematic method to assess convergent and discriminant validity is called a **multi-trait multi method matrix** approach. In this approach we use different instruments, for example different questionnaires or observation and self-report instruments to measure two traits.

Let’s take cat fondness and pizza fondness as an example. We would expect a very high association between cat fondness, measured observationally and measured through self-report. And the same goes for pizza fondness; we would expect the observational and self-report instruments of pizza fondness to show strong convergence.

We would expect a very small to zero association between cat fondness and pizza fondness both measured using self-report. A small association is possible because some people tend to give socially desirable or generally positive answers to questionnaires. The same zero to very small association can also be expected between cat and pizza fondness measured by observation. Finally, we would expect no association between different constructs measured with different methods.

If the relations show all the expected patterns then we have indirectly supported the construct validity of these four instruments. Of course this is a laborious process, because a lack of convergent or discriminant validity could be due to any one of the instruments. This would require a new study that combines the instruments with others in new ways to find out where the problem exactly lies. Hopefully you can appreciate how challenging it is to assess the construct validity of social and psychological constructs.

4.06 Measurement: Measurement reliability

A measurement instrument should be valid, but also reliable. **Measurement reliability** refers to the instrument's **consistency** or **stability** or **precision**. A reliable instrument will result in highly similar scores if we repeatedly measure a stable property in the same person.
My bathroom scale, for example, isn't *perfectly* reliable. If I step on it three times in a row it will usually show two or three different readings. But as long as the readings differ by one or two hundred grams, the scale's reliability is good enough for me.

So how do we determine the reliability of instruments that measure social and psychological constructs? Well in some cases it's as easy as administering the instrument twice to a group of participants and determining how strongly the results from the first and second measurement agree. This is called **test-retest reliability**. We can use this method if we're measuring things like weight or reaction times, but as soon as a person's memory of their answer is involved, things become more complicated.

Suppose I have a questionnaire measuring fondness of cats. It consists of five questions. If I ask a group of my friends to fill out this questionnaire once and then again fifteen minutes later, they will probably still remember the answers they gave the first time.
Now if I find a high consistency in the scores on the first and second measurement, is this because the instrument is reliable or because my friends have pretty good memories and would like to seem consistent in their attitudes?

One way to solve this problem is to look at the consistency, not between different times, but between different parts of the instrument at one time. This is referred to as **internal consistency**. We compare responses on the first three and the last two questions. Of course you can only do this if the instrument consists of several questions that are supposed to be comparable and measure the same construct.
If this is the case then you can determine the **split-halves reliability** by randomly splitting the test in half and assessing the association between the first and second half. There are also statistics that are equivalent to the average of all possible ways to split of the test.

If measurement consists of observation instead of self-report, you can have the observer rate the same behavior twice and assess the association between the two moments. This is referred to as **intra-observer consistency** or reliability.

Of course the memory of the observer can inflate the association. Since it shouldn't matter who makes the observations you could also assess the reliability of observation by having two different people observe and rate the behavior and look at the association between the two raters' scores. We call this **inter-observer consistency** or **inter-rater reliability**.

Ok so we've seen different ways to establish how reliable, or precise an instrument is. But what is it that makes an instrument less reliable? If the instrument perfectly reflects someone's **'true score'** or true value on the property of interest, then the measurement result, or **'observed score**' should be the same every time.

But what if we systematically measure an additional construct? Take my cat fondness scale, what if these questions also tap into the construct 'general positive attitude'? This could result in a systematically higher score for people with a positive attitude. We call this **systematic error**. This means our instrument is less valid, but not less reliable.

As long as the observed score is determined only by the **'true score'** on cat fondness and the **'systematic error'** caused by the second construct, positive attitude, then we would still get the same observed score every time we measure the same person.

Reliability is influenced by **random error**, error that's entirely due to chance. If the observed score is in part determined by random fluctuations, then we get different values each time we measure the same person.
If a scale is entirely unreliable, if there is no association between observed scores at different measurement moments, then we're basically measuring random error or noise. Put another way, this means that at least some reliability is required before an instrument can be valid. The reverse does not hold. A perfectly reliable instrument can be entirely invalid. This happens when it perfectly measures a different construct than it was supposed to measure.

Let's consider the possibilities in more detail. Of course the worst-case scenario is when an instrument has low reliability and low validity: a lot of random and systematic error. Even if the true score contributes a little to the observed score, it will be almost impossible to distinguish this contribution.

An instrument can also have low reliability and high validity: a lot of random error but very little systematic error. We are measuring the right property, just very imprecisely. An instrument can also have high reliability and low validity: a small amount of random error but a lot of systematic error. We're measuring the wrong property very precisely.

Best-case scenario is high reliability and of course high validity: a small amount of random error and systematic error. The observed score is mainly

determined by the true score. We are measuring the right construct with great precision.

Of course the trick is to separate all these error components from the true score, even if there is a fair amount of systematic and random error. Psychometricians and sociometricians aim to do this by using statistical modeling to partial out the random and systematic error.

4.07 Measurement: Survey, questionnaire, test

Surveys, **questionnaires** and **tests** are very popular measurement instruments in the social sciences. **'Survey'** is a general term that can refer to a list of questions asking about biographical information, opinions, attitudes, traits, behavior, basically anything. Surveys generally cover a variety of topics.

The term **'questionnaire'** is used when the focus is on one construct, or a related set of constructs, usually psychological traits, emotional states or attitudes. The term **'test'** is used when the aim is to measure an **ability**, such as general intelligence or math proficiency.

Surveys, questionnaires and tests should always include a clear instruction. For example, for a math test, it's important to know whether you should choose the right answer or the best answer and to how many decimals numerical answers should be rounded. The instruction can also provide a cover story. This will prevent participants from trying to guess the purpose of the study, possibly distorting their responses.

Surveys can be administered by an interviewer who visits respondents, goes through the questions and records the answers. This is very expensive though, so a more common method is to use self-report. This means people read, and respond to the questions by themselves. The survey can be completed using paper-and-pencil but of course the use of online applications is becoming much more common. Online administration is easier for the respondent: no need to visit the post office, you can complete the survey in your own time and help buttons and pop-up comments can provide extra instruction if necessary.

Online administration offers *researchers* added control: control over the order in which questions are viewed, checks to ensure all required answers are actually filled in and identification of strange response patterns - like the same answer to every question. A disadvantage of online administration is the low response rate. People easily delete an email with a link to an online questionnaire. It's much harder to turn away an interviewer at your door!

Surveys, test and questionnaires all consist of a series of questions. We refer to the questions as **items**, because sometimes they consist of

statements or even single words that respondents can agree or disagree with. The question, statement or word that a participant has to respond to is called the **stem**. The stem is usually accompanied by a set of discrete **response options** or a continuous range to choose from.

A psychological attitude, trait or state is almost always measured with items that describe feelings, thoughts or behavior that represent the relevant property. Usually, several items are used to measure the same construct. By using more than one item, random errors will generally 'cancel out'. Using several items also allows us to assess reliability by checking the internal consistency of the items.

Suppose I want to measure fondness of cats with the following five items:

1. Petting a cat is enjoyable.
2. I hate it when a cat jumps on my lap.
3. When near a cat I'm afraid to get scratched.
4. I frequently look at cat videos on the Internet.
5. Twenty years from now I can see myself having a cat.

People can choose from three answer options: Disagree, neutral or agree, scored 1, 2 and 3.

Items that are supposed to measure the same construct, or the same aspect of a construct, form a **scale**. When the items that form a scale are added together we get a sum score that indicates a person's value on the property.

Of course in our example agreement with some items indicates high cat fondness while agreement with others indicates low cat fondness. The items that are negatively worded, items 2 and 3, need to be **recoded**. Disagreement should be coded as 3 and agreement as 1! After recoding, adding the item scores results in a possible scale score between five and fifteen. A higher sum score means someone is more of a cat person.

Questionnaires frequently measure different aspects or dimensions of a psychological property by using subscales. Different sets of items tap into different aspects or maybe even different but related constructs altogether.

For example, if I'm interested in measuring your academic motivation I could distinguish between intrinsic motivation, extrinsic motivation and fear of failure. There are statistical methods that assess whether these dimensions are in fact distinguishable based on the pattern of responses provided by the respondents of course.

4.08 Measurement: Scales and response options

The most commonly used type of scale is a **Likert scale** or **summative** scale. Such a scale consists of comparable statements that are supposed to measure the same property. Respondents can indicate to what extent they agree with or endorse each statement.

Likert items should be monotone, meaning that respondents are consistently more likely to agree with the item if they possess the property to a greater degree. This is necessary because the scores on the items will be added.[1] Let's take my cat fondness questionnaire as an example. Items one, four and five will show stronger agreement from people who are fond of cats; items two and three will show stronger disagreement. After items two and three are recoded, all items are monotone: A higher score indicates higher cat fondness.

An example of a *non*-monotone item is "Cats are a necessary evil to get rid of mice". People who love cats will disagree, but extreme cat haters will also disagree, they will feel cats are just evil and not necessary at all. A high score could indicate high or low cat fondness.

There are other types of scales, such as differential scales that allow for non-monotone items and cumulative scales where the items themselves should show consistent ordering, each item expressing the property more strongly than the previous one. These scales are not used very often though.

Likert items generally have three to seven discrete response options, indicating strength of agreement. But a **visual analog** or **graphic rating scale** can also be used. This is a graphic representation, usually a simple line segment with two extremes at each end, like "disagree" and "agree". A respondent simply marks the line to indicate their position.
Some items will be "harder" to endorse than others because they represent more extreme cat fondness. Psychometricians and sociometricians use statistical techniques such as **item response theory** to assess an item's "difficulty".

OK, so how are Likert scale questionnaires constructed? Well the construction of a questionnaire starts with a clear **description** of the property to be measured. Often different **dimensions** can be distinguished, which are measured with a subset of items that form a **subscale**.

---

[1] Items are:
1. Petting a cat is enjoyable.
2. I hate it when a cat jumps on my lap.
3. When near a cat I'm afraid to get scratched.
4. I frequently look at cat videos on the Internet.
5. Twenty years from now I can see myself having a cat.

Identifying dimensions requires an in-depth analysis of the construct. It helps to consider different types of situations or areas in which the property can be expressed. For example: academic motivation can be intrinsic - you love the subject matter - but it can also be extrinsic - you expect to get a better job if you graduate.

It also helps to consider the different ways in which a construct can be expressed. For example, academic motivation can be expressed in attitudes and behavior; aggression can be expressed verbally or physically.
Once the relevant dimensions are identified, items are generated for each unique combination of dimensions. Ideally, each item describes a specific situation and a specific expression of the property. Respondents can easily misinterpret vague or too general items.

In general, all items, not just Likert items, should be **well formulated.** This means item formulation should be **short** and **simple**. Things like **double negation**, **unfamiliar words** and overly **complicated formulations** should be avoided.
The item "I do not take an unfavorable stance toward physical interaction involving fur-stroking with creatures of the feline genus" will probably confuse many respondents.

Formulations should also be **unambiguous**. Take the item "Cats are funny creatures": Does this mean that cats make you laugh, or that you think they're strange? **Double-barrelled** questions are another source of ambiguity. Take the question "I love to pet and pick up cats". How should people respond who like to pet cats but don't like picking them up?

Also, items should be neutral, **not suggestive**. An item like "Don't you agree that most cats are friendly?" could influence impressionable people to give a more favorable answer than they would have given otherwise.

Items should be **answerable** for all respondents. For example, people who have never been in direct contact with a cat cannot answer items 1, 2 and 3 of my questionnaire. Perhaps I should have included a **filter** question, asking if people have ever physically interacted with a cat.

Extreme wording should also be avoided. Using words like 'never' or 'always' make items very hard to moderately agree or disagree with. Take the item "Cats are purely evil creatures". Even moderate cat haters will probably disagree with this statement.

Of course **response options** also need to be unambiguous and consistent. Response options need to be **exhaustive** and **mutually exclusive**; exhaustive means all respondents should be able to reflect their position on the property. If you ask people their age and only provide the options twenty to thirty, thirty to forty and forty to fifty, then teenagers and people over fifty cannot respond.

These categories are also not mutually exclusive; people who are thirty or forty will have a hard time choosing between categories. Changing the options to zero to thirty, thirty-one to forty and forty-one and over fixes both problems.

Of course there is much more to item, scale and questionnaire construction than the points I've discussed here. There are many more scales types, item types and response options, and different ways to implement these, for example by combining self-report with ratings by others. Also, there are many more aspects to consider, such as item order and how to deal with sensitive questions. We've only just scratched the surface here.

4.09 Measurement: response and rater bias

When people respond to items, our hope is that their observed scores are determined mostly by their 'true' score on the measured property: The score we would obtain if we had a perfectly valid and reliable instrument. Of course no instrument is perfectly valid or reliable. There's always some degree of random and systematic error. A special type of systematic error occurs when respondents show a systematic bias in their responses to items. These biases are referred to as **response sets** or **response styles**.

I'll first discuss the most common response styles or biases that occur in self-report measurement. These are acquiescence, social desirability, extreme response styles and bias towards the middle.

Acquiescence refers to the tendency to agree with all statements, regardless of their content. An easy way to spot this bias is to include some negatively phrased items. Consider my cat fondness questionnaire. Items two and three are negatively worded. Someone who agrees with statements one, four and five, but also with items two and three, isn't answering consistently.

Social desirability causes a biased pattern that is a little harder to detect. A social desirability bias affects the responses of people who tend to present themselves more favorably or in more socially acceptable ways. A social desirability bias can occur if a scale measures a property that is considered socially sensitive, or measures a property that is relevant to someone's self-image.

It's possible to detect a social desirability bias by adding 'social desirability' items such as: "I've never stolen anything in my life" or "I've never lied to anyone". The idea is that every one has stolen something or lied at least once in their lives, if only as a child stealing from or lying to their peers. If people strongly agree with these items there's a fair chance that their responses to other questions are biased towards responses that are more socially accepted.

An **extreme response style** is much harder to detect. This bias can occur for example when respondents don't want to think about exactly *how* strongly they agree or disagree with an item, they'll just choose the most extreme option. So unlike the

acquiescence bias, participants' responses are consistent, just more extreme than their true value.

Bias towards the middle is highly similar to an extreme response style; only the tendency here is to choose a *less* extreme response option. For example, some respondents might never use the most extreme response options because they want to seem nuanced.

The ultimate version of bias to the middle occurs when there is an uneven number of response options, and a respondent always chooses the middle option. This response pattern can be detected by including some extra extremely strong items, such as "cats are purely evil creatures".
Cat lovers will strongly disagree, but even people who like cats just a little should show some disagreement with this statement. If they respond with the middle category to all items, including these extremely worded items their response pattern is inconsistent.

Biases due to the mere act of responding to a test or questionnaire can also occur when a rater observes and rates behavior of others. There are many rater biases. I will just discuss the halo effect and generosity and severity errors here.

The **halo-effect** occurs when a positive or negative rating on one dimension spills over to other dimensions of behavior that are rated or evaluated. A well-known example is that more attractive people are generally rated as more intelligent or better at their job.

A **generosity error** or **leniency effect** occurs when the rater is overly positive or kind in their ratings. The opposite, a systematic bias towards very strict or negative rating is referred to as a **severity error.**

It can be hard to detect, let alone avoid halo-effects and generosity or severity errors. One approach is to use several raters that are trained to use clearly defined coding schemes. Checking the inter-rater reliability and average ratings for each rater can help to detect any systematic bias between raters, but of course it is very hard to detect bias that is shared by all raters!

4.10 Measurement: Other measurement types

I've focused on measurement using **questionnaires**, **surveys** and **tests**, but of course there are many other ways to measure social and psychological constructs. In biology, medicine and psychology **physical** measures are very common. Think of things like electrical skin conductance to measure arousal, eye tracking to measure focus of attention, EEG and fMRI to register brain activity and reaction times to assess cognitive ability.

Another way to measure is through observation, a method frequently used in sociology, psychology and educational sciences. **Observational measurement** might seem simple, but there is more to it than just observing and recording all the behavior that you see. Systematic observation involves careful registration of specific behavior.
Researchers employ coding schemes that specify categories of behavior and their criteria. They specify what the behavior in each category looks like, how long it should be displayed and under what circumstances it should occur.

A researcher also needs to decide on the time frame to be coded. Will we view an entire hour of videotaped behavior or will we sample five two-minute intervals? If we have taped material of an hour for each of sixty participants, then the two-minute intervals might be a better idea!
Other important issues are training and calibration of observers. Coding schemes can be complex and target behavior can be difficult to spot. So it's a good idea to have more than one observer and to train observers until they show enough agreement when coding the same material.
Agreement between different observers, called the inter-rater reliability, should be high of course. If reliability is low than at least one of the observers codes the behavior differently from the rest. Or, even worse, the behavior cannot be interpreted consistently.

Let's move on to a related form of measurement. **Trace measurement** assesses behavior indirectly through physical trace evidence. An example is counting the number of used tissues after a therapy session to represent how depressed a client is. Sometimes a property can be represented with measurements that were already collected by others. We refer to this as **archival data**. An example of archival research is to use census data, collected by a national research institute, on income and voting behavior. We could investigate whether areas with a higher average income are associated with more votes for conservative political parties. Trace measurement and especially archival data are frequently used in political sciences and sociology.

Content analysis is a technique that shares characteristics with observational, archival and trace measurement. Like observational measurement, content analysis consists of structured coding, but of elements in a text. The text can consist of newspaper articles, blogs, narratives or transcription of interviews.

Content analysis can be used, for example, to see if conservative and liberal politicians argue differently, by identifying the number of emotional and rational words they use in newspaper interviews. Of course this is a simple example. Text can be coded automatically according to very complex schemes using computer software.

A final measurement method I want to discuss is interviewing. In a **structured interview,** the questions, the question order and response options are pre-determined. This type of interview - be it face-to-face, through telephone or Skype - is not much different from using a survey. The response rate of interviews is higher, but it can be more difficult to get unbiased answers to sensitive questions.

Unstructured or **open interviews** are very different. An open interview is considered a qualitative method. Now although the focus here is on quantitative methods, I quickly describe open interviews because they are used very often. In an open interview the interviewer starts off with a general topic and usually has a set of points to be addressed but the interview is not limited to these points.

Questions are open-ended and there is little structure, so the conversation can lead anywhere depending on the respondent's answers. The questions that will come up and the range of answers are undetermined. Of course this makes it much harder to compare and aggregate data from different respondents. I won't go into other qualitative methods here, but you should know there are other methods available such as case studies, focus groups, oral histories, participatory observation and many, many more.

Module 5: Sampling

5.01 Sampling: External validity threats

External validity, or generalizability, refers to whether the hypothesized relation holds for other persons, settings and times. Just like with internal validity, there are a number of threats to external validity.

A **history** threat means that the observed effect doesn't generalize to other time periods. Consider a compliance study performed in the nineteen fifties in the US. Results showed that participants were willing to comply with highly unethical directions provided by an authoritarian experimenter. These results would probably be less extreme if we repeated the study nowadays, for example because people are more highly educated and less sensitive to authority than in the nineteen fifties.

A **setting** threat to external validity means that the observed effect only holds in a specific setting. In other words, the findings do not generalize to other environments or situations. Suppose we investigate the relation between violent imagery and aggression and find that children who watch a violent video are more aggressive afterwards in the school playground. A setting threat occurs if this effect depends on the surroundings, for example if children are *not* more aggressive when they play at home under their caregiver's supervision.

There are two setting threats associated with the ***artificiality*** of the *research* setting specifically. These threats are **pretesting** and ***reactivity***. A ***pretesting*** threat means that the observed effect is found only when a pretest is performed. This threat is closely related to the ***internal validity threat*** of ***testing***.
Say we investigate a new therapy for treating depression and use a pretest. Suppose the depression pretest makes participants realize how serious their problems are, and thereby makes them more receptive to the treatment. The treatment is effective, but only if receptiveness is increased by the pretest first. In this case internal validity is threatened because 'receptiveness' is missing from our hypothesis. External validity is also threatened, because the hypothesis will only apply to situations where a pretest is part of the setting.

The second artificiality threat is **reactivity**. A reactivity threat occurs when the participants or experimenter react to the fact that they are participating in a research study. Reactivity includes participant and experimenter expectancy and altered participant behavior, for example due to nervousness. This can cause the hypothesized relation to occur only in a research setting and not in a in a natural setting. Say we investigate a new method for teaching high school math. The researcher is present during the lessons and measures math performance in class. What if students work harder because they know they are being studied and this makes the new method more effective? In a natural setting, without the researcher present, students might put less

effort in their schoolwork, reducing the effectiveness of the new method.

Selection is a final and very important threat to external validity. A selection threat occurs when the hypothesized relation only holds for a specific subset of people or if the results in our study are biased due to over- or underrepresentation of a certain subset.
Suppose that in our study on a new depression therapy, we recruited participants who actively volunteered. Say we find that the therapy method is effective. Of course our volunteers might be more proactive about solving their problems than the average person. It is entirely possible that the method is *in*effective for people who are less proactive. The overrepresentation of volunteers might lead to an overestimation of the therapy's effectiveness.
Another example: Suppose we want to know people's opinion on women's right to vote and we interview people on a university campus. The sample is now so selective that it is highly unlikely that results will generalize to the general public's opinion.

What can we do about these threats to external validity? Well history and setting threats to external validity can be reduced by replicating a study in a different time or by repeating a study in different settings. In the case of threats related to the artificiality of the *research* setting specifically, this means repeating a study in a more natural environment. Replication can also reduce the threat of *selection* to external validity, in this case by repeating a study with different groups of subjects. But there is another way to reduce the threat of selection. I'm referring to **random sampling** of the research sample, also referred to as **probability sampling**.

5.02 Sampling: Sampling concepts

Some sampling methods offer better protection against the selection threat to external validity than others do. To understand why, you first need to be familiar with some basic sampling concepts. The two most important concepts are ***population*** and ***sample***.

The selection threat concerns the generalization of findings to other persons. Exactly what other persons are we referring to? All people in the entire world, or just people in our country or culture? Researchers should anticipate this question by defining their target ***population*** explicitly. The term **population** refers to the entire collection of people or groups to whom the hypothesis is supposed to apply.

Let's look at two examples of populations. Consider the hypothesis "Loneliness causes an increase in depression". This is a typical universalistic hypothesis. If the population is not explicitly mentioned, we infer the relation is assumed to hold for all people, in all cultures, in the past, now and in the future.
Another example: "Patriotism is steadily declining in the Netherlands over the last five years". This is at typical particularistic hypothesis. It is clear that this hypothesis applies to a specific country and to a specific time.

Let's assume for a minute that the target population for a hypothesis is clearly defined. How can we determine if the results generalize to this entire population? Well if we measure the entire population, then we're automatically sure the results hold for the entire population, everybody was measured.
Of course for universal hypotheses it's simply impossible to measure the entire population, because the population consists of all people everywhere, including all people who are long dead and all people who have yet to be born.
Even if the target population is smaller and well defined, it is almost always too complicated and too expensive to include the entire population in a study. This is why we take a **sample**: a subset of the population. The sample is used to represent or estimate a property of the population.

Of course it's possible that this sample does not represent the population accurately. Suppose we sample mostly elderly people in our study on the effect of loneliness on depression and we find a strong effect. The overrepresentation of a specific part of the population can weaken the study's external validity. Perhaps the strong effect of loneliness on depression is less apparent for young people. If our sample had been more representative of the entire population we would have found a smaller effect.
The same goes for our study of decreased patriotism. Suppose our sample consisted mainly of highly educated people working at a university. This might lead us to underestimate patriotic attitudes in the Netherlands. Our results will be biased.

We will consider different sampling methods and see how they deal with the selection threat to external validity. But before we can do so, there are some terms you need to become familiar with.

An **element**, or unit, is a single entity in the population. Together all the elements form the population. An element most often consists of one person, but of course it depends on your hypothesis. An element can also be a group, a school, a city, a union, a country; you name it.

A **stratum** is a subset of elements from the population that share a characteristic. In the population of currently enrolled students from the University of Amsterdam we can distinguish a female and a male stratum, for example. Of course we can identify many different strata that may overlap, for example male and female undergraduate and graduate students.

The term **census** refers to an enumeration or *count* of all elements in the population. The term can also refer to a situation where all elements in the population are actually measured; in that case, the sample consists of the entire population. The term census can indicate a 'national census': A nation-wide survey where demographic information on each inhabitant is collected. Of course in many western countries this census is conducted virtually, by collecting information from government databases.

A final term that you need to be familiar with is the term **sampling frame**. A sampling frame is essentially a list of all the elements in a population that

can be individually identified. A sampling frame can overlap with a census defined as an enumeration of the population. A sampling frame is more than a simple list of elements however.

A sampling frame provides a way of actually contacting elements. It could be a phonebook or a list of email addresses for all students currently enrolled at the University of Amsterdam, for example. Also, a sampling frame doesn't always include all elements of a population. This could be due to clerical errors or an outdated list. Ok, you now know the basic concepts necessary to learn about different sampling methods.

5.03 Sampling: Probability sampling

Probability sampling minimizes the selection threat to external validity. Before I discuss different types of probability sampling, let's consider the essential feature of probability sampling and how this feature helps to minimize the risk of systematic bias in our selection of participants.

The essential feature of probability sampling is that for each element in the sampling frame, the probability of being included in the sample is known and non-zero. In other words, some form of random selection is required where any element could in principle end up in the sample. To use probability sampling, we need to have a sampling frame: a list of all elements in the population that can be accessed or contacted. A sampling frame is necessary to determine each element's probability of being selected.

Now let's see why probability sampling minimizes the threat of a systematic bias in our selection of participants. Reducing systematic bias means reducing the risk of over- or underrepresentation of any population subgroup with a systematically higher or lower value on the property. Otherwise, our sample value will be unlikely to represent the population value accurately.

We've already seen a method to eliminate systematic bias in participant characteristics. Remember how we eliminated the ***selection threat to internal validity***? We used ***random assignment*** to get rid of systematic differences between the experimental and control condition. In the long run any specific participant characteristic will be divided equally over the two groups. This means that any characteristic associated with a systematically higher or lower score on the dependent variable cannot bias the results in the long run.

The same principle can be applied, not in the *assignment*, but in the **selection** of participants. We avoid a systematic difference between the sample and the population, by randomly selecting elements from the population. In the long run any specific participant characteristics will be represented in the sample,

proportionally to their presence in the population. We call this a **representative sample**.
Suppose a population consists of eighty percent women. With repeated random sampling, we can expect the sample to contain eighty percent women in the long run. Each individual element has the same probability to be selected, and since there are more women, female elements will be selected more often.

Besides resulting in a ***representative sample*** in the long run, probability sample has another advantage. Probability sampling allows us to assess the **accuracy of our sample estimate**. Probability sampling allows us to determine, that with repeated sampling, in a certain percent of the samples, the sample value will differ from the real, population value by no more than a certain **margin of error**.

This sounds - and is - complicated. But it basically means that we can judge how accurate our sample estimate is in the long run. Given a certain risk to get it wrong, we can assess what the margin of error is on average, meaning by how much the sample and population value will differ on average.

Consider an election between conservative candidate A and democratic candidate B. We want to estimate the proportion of people in the population that will vote for candidate A as accurately as possible. Random sampling allows us to make statement such as this: If we were to sample voters repeatedly, then in ninety percent of the samples, the true, population proportion of votes for A would lie within eight percentage points of our sample estimate.
So if we find that sixty percent of our sample indicates they will vote for A, then we can say we are fairly confident that the true proportion will lie somewhere between fifty-two and sixty-eight percent. This interval is called a **confidence interval**. Of course this particular interval could be wrong, because in ten percent of the samples the sample value will lie further that eight percentage points from the true value. This could be one of those samples, so we can never be certain.

5.04 Sampling: Probability sampling - simple

There are several types of probability sampling. In this video I'll discuss the two simplest types: ***simple random sampling*** and ***systematic sampling***.

The most basic form of probability sampling is **simple random sampling**. In simple random sampling each element in the sampling frame has an equal and independent probability of being included in the sample. Independent means the selection of any single element does not depend on another element being selected first. In other words, every possible combination of elements is equally likely to be sampled.

To obtain a simple random sample, we could write every unique combination of sampled elements on a separate card, shuffling the cards and

then blindly drawing one card. Of course, if the population is large, then writing out all possible combinations is just too much work.
Fortunately, an equivalent method is to randomly select individual elements. This can be done using random number tables, still found in the back of some statistics books. But these tables have become obsolete; we can now generate random number sequences with a computer. For example, if our population consists of twelve million registered taxpayers, then we can generate a sequence of two hundred unique random numbers between one and twelve million.

Systematic sampling is a related method, aimed to obtain a random sample. In systematic sampling only the first element is selected using a random number, the other elements are selected by systematically skipping a certain number of elements.
Suppose we want to sample the quality of cat food on an assembly line. A random number gives us a starting point: say the seventh bag. We then sample each tenth bag, so we select bag number seven, seventeen, twenty-seven, etcetera. It would be much harder to select elements according to random numbers, say bag number seven, thirty, thirty-six, forty-one etcetera., especially if the assembly line moves very fast.
With this approach, each element has an equal probability of being selected, but the probabilities are not independent. Elements seventeen, twenty-seven, thirty-seven etcetera, are only chosen if seven is chosen as a starting point. This is not a real problem; it just requires a little more statistical work to determine things like the margin of error.

The real problem with systematic sampling is that it only results in truly random sample if there is absolutely no pattern in the list of elements. What if the assembly line alternately produces cat food made with fish and cat food made with beef? Let's say all odd-numbered elements are made with fish. In our example we would never sample the quality of cat food made with beef!
Of course this is an exaggerated example, but it illustrates that systematic sampling can be dangerous. A pre-existing list or ordering of elements can always contain a pattern that we are unaware of, resulting in a biased sample.

So systematic sampling only results in a truly random sample if it is absolutely certain that the list of elements is ordered randomly. We can make sure of this by randomly reordering the entire list. We could generate a sequence of random numbers of the same size as the list and then select elements from this list using systematic sampling.
Of course this is equivalent to random selection directly from the original list using random numbers. Unless we can be sure that the list is truly random, systematic sampling should *not* be considered a form of probability sampling; instead it should be considered a form of non-probability sampling.

5.05 Sampling: Probability sampling - complex

There are many sophisticated probability-sampling methods. I'll discuss two methods that go beyond the basic idea of random sampling, but are still relatively simple. These are **stratified random sampling** and **multi-stage cluster sampling**.

In **stratified random sampling** we divide the population into mutually exclusive strata. We sample from each stratum separately using ***simple random sampling***. The separately sampled elements are added together to form the final sample. Stratified random sampling is useful for two reasons.

First, it allows us to ensure that at least in terms of the sampled strata, our sample is ***representative***. This means subpopulations are represented in the sample in exactly the same proportion as in the population. With simple random sampling we can expect the sample to be representative in the long run, but due to chance, in any particular sample, strata might be over- or underrepresented.
Second, stratification is useful because it can make sampling more efficient. This means, all other things being equal, that we achieve a smaller margin of error with the same sample size. Stratifying only increases efficiency if the strata differ strongly from each other, relative to the differences within each stratum.

Imagine we want to sample the quality of cat food produced on an assembly line. The line produces cat food made with fish and cat food made with beef. Suppose the average quality of beef cat food is higher than that of fish cat food. Also, the quality varies relatively little when we consider each type of food separately. Under these circumstances we will obtain a more accurate estimate of the population's average food quality if we stratify on food type.
This is because quality is related to food type; even a small overrepresentation of one food type can distort our overall estimate of food quality. Stratifying prevents this distortion. If the quality does not differ between food types, then overrepresentation of one food type will not distort the overall estimate and stratification will not improve efficiency.

It is important to realize that stratified sampling requires that we know which stratum each element belongs to. If we can *identify* strata, then we also know their size. As a consequence, the size of our subsamples does not have to correspond to the size of the strata. We can calculate a representative estimate by weighing the subsamples according to stratum size.
Why would we do this? Well, suppose our stratum of fish cat food is relatively small, or is known to strongly vary in quality. In both cases our estimate of the quality of fish cat food might be much less likely to be accurate than that of beef cat food. It might be worth it to take a bigger sample of fish cat food, so we have a better chance of getting an accurate estimate. Of course this means over-representing fish cat food.

We can correct for this overrepresentation by weighing the separate estimates of fish and beef cat food according to their stratum sizes before

averaging them into an overall estimate of food quality. This way the sample value is representative, efficient and more likely to be accurate.

Let's turn to multi-stage cluster sampling, the final type of random sampling I want to discuss. **Multi-stage cluster sampling** allows us to use random sampling without going bankrupt. Consider sampling frames that consist of all inhabitants, students, or eligible voters in a certain country. If we were to randomly select elements from these frames we would have to travel all over the country. In most cases this is just too expensive.

A solution is to randomly sample in stages, by first selecting clusters of elements. Say we want to sample math performance in the population of all Dutch students currently in their third year of secondary education. We start by forming a sampling frame of all school districts; this is the first stage, where students are clustered in districts. We randomly select a very small sample of school districts. We can use stratification to make sure we include districts in urban and rural areas.
In the second stage we randomly select schools from the previously selected districts. Students are now clustered in schools. In the third stage third year math classes are randomly sampled from the previously selected schools. We could even include a fourth stage where students are randomly sampled from the previously selected classes. Stratification can be used in all of these stages.

Multi-stage cluster sampling makes random sampling feasible. But the margin of error is harder to determine, because the probability to be included in the sample is no longer the same for all elements, like it was with simple random sampling. Also, cluster sampling is usually associated with a larger margin of error, even if stratified sampling is used to increase efficiency. However, these disadvantages are generally more than outweighed by the reduction in cost and effort.

5.06 Sampling: Non-probability sampling

Probability sampling can be contrasted with **non-probability sampling**. In non-probability sampling some elements in the sampling frame either have zero probability to be selected or their probability is unknown. As a consequence, we cannot accurately determine the margin of error. It's also impossible to determine the likelihood that a sample is representative of the population.
There are several types of non-probability sampling. I'll discuss the four most common types: **convenience sampling**, **snowball sampling**, **purposive sampling** and **quota sampling**.

Convenience sampling, or ***accidental sampling***, is the simplest form of non-probability sampling. In convenience sampling, elements are selected that are the most convenient, the most easily accessible. For example, if I'm interested in investigating the effectiveness of online lectures on study performance, I could recruit students in courses that I teach myself. Of course this a highly selective

sample of students from a particular university in a particular bachelor program. Results will almost certainly be influenced by specific characteristics of this group and might very well fail to generalize to all university students in my country, let alone students in other countries.

So the risk of bias is high and we have no way to determine how closely the sample value is likely to approach the population value. Even so, convenience samples are used very often. Because sometimes, it's simply impossible to obtain a sampling frame. In other cases, the effort and expense necessary to obtain a sampling frame are just not worth it; for example when a universalistic, causal hypothesis is investigated.

Snowball sampling is a specific type of convenience sampling. In snowball sampling, initially, a small group of participants is recruited. The sample is extended by asking the initial participants to provide contact information for possible new participants. These new participants are also asked to supply contacts. If all participants refer new ones, the initially small sample can grow large very quickly.

Suppose we want to sample patients who suffer from a rare type of cancer. We could approach a patient interest group, for example, and ask the initial participants if they can put us in contact with other patients that they know through other interest groups or through their hospital visits. We continue to ask new participants to refer others to us, until the required sample size is reached.

Snowball sampling is very useful for hard-to-reach, closed-community populations. Of course all disadvantages of convenience sampling also apply to snowball sampling, maybe even more so, because there is the added risk that we are selecting a clique of friends, colleagues or acquaintances. These people could share characteristics that differ systematically from others in the population.

In **purposive sampling,** elements are specifically chosen based on the judgment of the researcher. A purposive sample can consist of elements that are judged to be typical for the population, so that only a few elements are needed to estimate the population value. A purposive sample can consist of only extreme elements, for example, to get an idea of the effectiveness of social workers working with extremely uncooperative problem families.

Elements can also be purposively chosen because they are very much alike, or reversely, very different, for example, to get an idea of the range of values in the population. Or, elements can consist of people who are judged to be experts, for example when research concerns opinions on matters that require special knowledge.
Purposive sampling is used mostly in qualitative research, so I won't go into further details here. Suffice it to say that purposive sampling suffers all the same disadvantages that convenience sampling does. The researcher's judgments can even form an additional source of bias.

Quota sampling is superficially similar to stratified random sampling. Participants in the sample are distinguished according to characteristics, such as

gender, age, ethnicity or educational level. The relative size of each category in the population is obtained from a national statistics institute, for example.
This information is used to calculate how many participants are needed in each category. So that the relative category size in the sample corresponds to the category size in the population. But instead of randomly selecting elements from each stratum, participants for each category are selected using convenience sampling. Elements are sampled until the quotas in all categories are met.

Although this approach might seem to result in a representative sample, all kinds of biases could be present. Suppose the choice of participants is left to an interviewer. Then it's possible that only people who seem friendly and cooperative are selected. If a study uses non-probability sampling, the results should always be interpreted with great caution and generalized only with very great reservation.

5.07 Sampling: Sampling error

The goal of sampling is to estimate a value in the population as accurately as possible. But even if we use the most advanced sampling methods, there will always be some discrepancy between our sample value - the estimate - and the true value in the population. The difference between sample and population value is generally referred to as error. This error can be categorized into two general types ***sampling error*** and ***non-sampling error***. In this video I'll only discuss the first type: sampling error.

It's important to keep in mind that the true value in the population is almost always unknown. If we knew the population value then we wouldn't need a sample. This also means that for any particular sample we cannot assess how large the error is exactly.

However, for ***sampling error*** it is relatively easy to estimate how large the error is. Let's look at sampling error in more detail and see how it works. If we would take an infinite number of samples from a population, then under certain conditions, the average sample value of all these samples will correspond to the population value.
But of course individual samples will result in sample values that are different from the population value. **Sampling error** is the difference between sample and population value that we would expect due to chance. We can estimate how large the sampling error is on average, if we were to repeatedly draw new samples from the same population. Note that this only works for randomly selected samples!

The average error, called the ***standard error***, can be estimated based on the values obtained in a single sample. We can then use the standard error to calculate a **margin of error.** You might think the margin of error tells us by how much our sample differs from the population at most. But we can't calculate between what boundaries the true population value lies exactly, because we are estimating the sampling error in the long run, over repeated samples. In the long run a ridiculously small or large value is always

possible. What we *can* say is that the population value will lie between certain boundaries *most* of the time.

This information is captured in a **confidence interval**. A confidence interval allows us to say that with repeated sampling, in a certain percentage of these samples, the true population value will differ from the sample value by no more than the **margin of error**.

Suppose we want to estimate the proportion of people that will vote for candidate A in an election. We sample one hundred eligible voters and find that sixty percent of the sample says they'll vote for A. We have to decide how confident we want to be. Let's say that with repeated sampling, we want the population value to fall within the margin of error at least ninety percent of the time. With this decision, we can now calculate the margin of error. Let's say that the margin of error is eight percent. This means we can say that with repeated sampling, the population value will differ from the sample value by no more than eight percent, in ninety percent of the samples.

Sampling error is related to the sample size. As sample size increases, sampling error will become smaller. Sampling error is also influenced by the amount of variation in the population. If a population varies widely on the property of interest, then the sample value can also assume very different values. For a given sample size, sampling error will be larger in a population that shows more variation.

Ok, so to summarize: sampling error is the difference between population and sample value due to chance, due to the fact that our sample is a limited, incomplete subset of the population. Sampling error is unsystematic, random error. It is comparable to the random error that makes a measurement instrument less reliable.
We can estimate how large the sampling error will be in the long run, which allows us to conclude how accurate our sample value is likely to be. This only works under certain conditions. One of these conditions is that the sample is a random sample from the population.

5.08 Sampling: Non-sampling error

Sampling error can be contrasted with non-sampling error. ***Sampling error*** is the difference between population and sample value due to the fact that our sample is a limited, incomplete subset of the population. ***Non-sampling error*** is the difference between population and sample value due to sources *other* that sampling error. Two major sources of non-sampling error are ***sampling bias*** and error due to ***non-response***. They are both related to the **sampling procedure**.

Sampling bias is a systematic form of error. Sampling bias is the difference between sample and population value due to a systematic under- or overrepresentation of certain elements in the population. Sampling bias occurs when

some elements have a much smaller or larger chance to be selected than was intended. Sampling bias can also occur when certain elements have no chance to be selected at all.
Suppose we want to estimate the proportion of people that will vote for candidate A in an election. Sampling bias could occur if participants were recruited on the street by an interviewer during working hours. This could lead to an underrepresentation of people who are employed full-time. If these people would vote for candidate A more often, then we would systematically underestimate the percentage of votes for candidate A.

The risk of sampling bias is eliminated, at least in the long run by using a probability sampling method. With non-probability sampling, the risk of sampling bias is strong. Sampling bias is comparable to the systematic error that makes a measurement instrument less valid, or less accurate.

Non-response is another source of error. Non-response refers to a lack of response to invitations or the explicit refusal to participate in a study. Non-response also includes participants who drop out during the study or participants whose data are invalid because they did not participate seriously, because something went wrong or they did not understand or failed to comply with some aspect of the procedure.
If non-response is random, then you could say that non-response results in a smaller sample and will thereby slightly increase the margin of error. But sometimes non-response is not random. Sometimes specific subgroups in the population are less likely to participate. If this subgroup has systematically different values on the property of interest, then non-response is a source of systematic error.
Suppose people with a lower social economic status are less likely to participate in polls and also prefer other candidates to candidate A. In that case we are missing responses of people that would not vote for A, which could lead to a systematic overestimation of the percentage of people that will vote for A.

Besides sampling bias and non-response, there are other sources of non-sampling error related to the sampling procedure. One example is an incomplete or inaccurate sampling frame, for example because the frame is out of date.

Apart from non-sampling error related to the sampling procedure, there are two other general types of non-sampling error.

The first type is error related to the **collection of data**. This type of error could be caused by errors in the instrument, such as poorly worded questions or untrained observers. Data collection errors can also be due to the errors in the procedure, such as giving inaccurate instructions, a failure of equipment or distraction by fellow participants during data collection.

A final source of non-sampling error lies in the **processing of data** after they have been collected. Data entry errors can be made for example when data is entered into a data file manually, or when responses need to be recoded or aggregated in the data file.

As you can see, non-sampling error includes systematic error such as sampling bias, systematic non-response error and systematic collection error due to faulty instruments or procedures. However, non-sampling error also includes random error, such as random non-response error, random data collection and random data processing errors.

One final remark: For random samples the sampling error can be estimated. The size of non-sampling error is much harder to assess, even in random samples. There are all kinds of sophisticated techniques available to assess sample bias and systematic and random non-response errors. Unfortunately, these techniques usually require rigorous sampling methods, large sample sizes and all kinds of additional assumptions.

5.09 Sampling: Sample size

The goal of sampling is to obtain the best possible estimate of a population value, within the limits of our budget and our time. Suppose we've decided on a sampling method for our study - preferably a probability sampling method if this is at all possible. The question now remains how many elements we need to sample in order to get an accurate estimate of the population value.

An easy answer would be "as large a sample as we can afford". Because as sample size increases, the margin of error will decrease. Accidental over- or underrepresentation of certain elements will be less extreme and will become less likely. In other words, a bigger sample is always better in terms of accuracy.

But this doesn't mean we should all collect samples consisting of tens of thousands of elements. This is because as the sample size grows, the decrease in the margin of error becomes smaller and smaller. At a certain point the cost of collecting more elements outweighs the decrease in the margin of error.
Say we want to estimate the proportion of votes for candidate A in upcoming elections. Suppose we have a sample of five hundred eligible voters. Then the error won't be cut in half if we double the sample to a thousand elements, the decrease in error will be much, much smaller.

Note that it is the absolute size of the sample that matters, not the relative size. It doesn't matter if we are estimating election results in Amsterdam, with slightly more than half a million eligible voters, or national elections with more than 13 million voters. As long as the samples are both randomly selected, the margin of error will be the same, all other things being equal. This seems very counter-intuitive, but it's true nonetheless.

Of course there are other factors to consider when deciding on sample size. The variability of the population is an important factor. Heterogeneity, or strong variation in the population on the property of interest, results in a larger margin of error, all other things being equal. If values in the population vary widely, then

a sample is more likely to accidentally over- or underestimate the true population value.

If the population is more homogeneous or similar, meaning it takes on a narrow, limited set of values, well then the sample value will automatically lie close to the population value. If a population is more homogeneous, we can sample more efficiently. This means, all other things being equal, that we can achieve a smaller margin of error with the same sample size. Or, conversely, we can obtain the same margin of error with a smaller sample.

If a probability sampling method is used we can determine what margin of error we are willing to accept, given a certain confidence level. We can say that we want our sample estimate of election results to differ by no more than five percent from the final results in ninety five percent of the samples, if we were to sample repeatedly.

We, or rather a computer, can now calculate exactly what sample size we need, to obtain this margin of error at this confidence level. This does require that we use random sampling and that we can estimate the variability in the population, for example based on previous studies, old census data or just a best guess if necessary.

I'll just mention one other important factor to consider when determining the sample size. It's a good idea to plan ahead and compensate for **non-response**. Non-response refers to elements in the sample that cannot be contacted, that refuse to participate, fail to complete the study or provide invalid responses.
If the response-rate can be estimated based on previous or comparable research, then we can take non-response into account and sample extra elements that will compensate for the expected loss of elements due to non-response.

Module 6: Practice, Ethics & Integrity

6.01 Practice, Ethics & Integrity: Documentation

In this video I'll discuss the relevance of research documentation. Good documentation is critical for two things: objective replication and checking and verification of results and conclusions. In other words documentation is very important to research integrity.

In order to replicate a study, the hypothesis, research design and predictions need to be clearly stated. Replication also requires a clear description of the procedures and instruments and the protocol for contacting and interacting with participants. Finally, the data, data manipulation and statistics that were performed and the researchers interpretation of these results into a final conclusion also need to be documented. If any of these steps are unclear or cannot be verified after the study is performed, then this can lead to confusion when contradictory results are found in replication studies.

The hypothesis, **research design**, specific predictions and interpretation of results are generally explicitly stated in a research publication or journal article. Such a document also contains information on the research procedure, instruments, data, data manipulations and statistics, but this information is always *summarized*.
Giving all the details simply takes up too much space. Most researchers just want to know a study's goal, setup and outcome. However, the detailed information should be documented somewhere, so it can be made available to those researchers who *are* interested in performing an exact replication.

So what details should be documented exactly? Well to start, all information on the **instruments** and **materials** need to be available. With instrument information I mean the instruments used to measure or manipulate the variables of interest and the procedures or instruments used to control or measure extraneous variables.
The materials include written or verbal instructions, texts used in email communications, consent forms, documents containing debriefing information, etcetera, etcetera. Basically all materials, information and documents that were used by the experimenter or participant need to be retrievable.

Let's focus on the instruments for a moment. Unlike tape measures and weight scales, social and psychological measures are usually not self-evident. In research articles measurement information is summarized by stating of how many questions a scale in a questionnaire consists, what the response options are, what the range of possible scale scores is. An example item is often provided.
This is the minimum required information to interpret the results, but a full version of the instrument needs to available for checking and replication. The instrument itself, meaning its observation coding scheme or questionnaire items and response options should be available to others.

This requires for example recording what version of a standardized test was used, or documenting and providing the instrument itself, if it's not publicly or commercially available.

Besides information about instruments and materials the **research protocol** needs to be documented. The research protocol refers to the order and manner in which materials and information were presented to participants and the way procedures were implemented. A research protocol is a clear description of what happens to the participant from start to finish. This includes the moment of first contact, how a participant is recruited, with what information and by whom. The protocol also describes how consent is obtained, what instruction or information is given in what order, what help is provided if participants don't understand instructions or they behave in an unexpected manner. The research protocol also describes how participants are debriefed about the research after their participation.

Documenting materials is relatively easy, it simply means saving emails and documents. Writing a research protocol is more tedious work and is not always performed diligently. Of course this can become a big problem when other researchers want to replicate a study and can only use summarized information or have to rely on the original researcher's memory!

Another thing that is often badly documented or not even explicitly formulated at all is the **statistical plan**. Ideally a researcher should determine what statistical analysis will be performed *before* looking at the data. This is because the choice of statistics can sometimes determine whether the predictions are confirmed or disconfirmed.
Ideally the hypothesis and intended statistical analyses are **preregistered**, so that a researcher is not tempted to change either after seeing the results and realizing that a different approach would result in more favorable outcomes. Preregistration is customary in medical sciences and is gaining popularity in neuroscience, clinical psychology and other fields.

Finally, information about pilot studies often fails to be documented. A pilot study is a small, preliminary study where a newly developed instrument or manipulation procedure is tested. Such studies can contain useful information for other researchers developing or adapting similar instruments. A pilot study should therefore at least be mentioned in the published article, so that others are aware that the results are available.

6.02 Practice, Ethics & Integrity: Data management

After the data are collected, a researcher's responsibility to be transparent and to document actions and decisions does not stop. After data collection, data need to be stored, processed and analyzed using statistics. Together this is referred to as **data management**.

Once the data are collected, they need to be recorded. Measurements are stored in a computer file, usually in a data matrix where columns represent variables and rows represent participants. Such a data file is useless if it's unclear what the recorded data represent. If we have a column representing sex for example, but we don't know whether a zero means female or male, the information is useless.

Information or meta-data about what the variables mean, is stored in a **codebook**. The codebook specifies what property each variable measures and what the values mean, what the range of possible values is and what values are used to denote a participant did not supply relevant information, referred to as missing values.

Suppose we collect responses to ten items forming a depression questionnaire with three answer options. To correctly interpret these scores we need to know that for each item the minimum possible value is one, the maximum value is three and we denote a missing response with the value nine. Without this information we would not realize that something went wrong if we find a score of five. We could have made for example an error entering the raw data into the computer.

Because data entry errors are always made it is extremely important to always save the original data. With the original data, I mean the paper questionnaires filled out by participants, or the video material used for observational coding. Without the original data we cannot check whether the data were entered into the file correctly. When data are entered manually it's always a good idea to let someone else enter a random selection of the data again and check for consistency. If it turns out the original entry is inconsistent then all the data need to be entered again, but more carefully of course!

The original data and instrument information are also necessary to check the codebook. Sometimes a codebook can be confusing or seem to be wrong. For example when responses to an item are unexpectedly low, this could be a valid pattern, but it could also be an error in the codebook. It is possible for example the codebook wrongly indicates an item is positively worded, when it is in fact phrased negatively and therefor should be recoded.

The original data file is also very important and should be stored separately. With the original data file I mean the file that contains the raw data as they were entered originally, before any manipulation was performed. Data manipulation refers to computations such as recoding and computing aggregate scores like a sum score of depression. Another example is calculating age from date of birth.

Without the original data file, we cannot check whether we made any errors in manipulating the data. Suppose that, for a negatively worded depression item, I change the score of one to a score of three - and three to one - and then I accidently recode again changing threes back into ones and ones into threes. I end up with negatively scored items, without being aware of this, thinking they are scored positively. If I find unexpected results, I can simply check if I made a recoding error by comparing against the original data file, that's why it's important.

Not only should we record the original data and data file, we should also record any processing, selection or computations we perform. Otherwise we might not be able to reproduce processed data that are used to formulate the final conclusions.
For example when I select a subset of my sample, say only people who completed the depression questionnaire within a month, then my results might be different from results obtained from the entire sample. If I don't record my selection criteria, then in a year from now I will probably have forgotten the exact criteria and will not be able to reproduce my own results.

Both the processing of data, for example recoding and computing sum scores, selection of data and the statistical analyses are generally recorded in a syntax file. The syntax file is like a simple programming file that can be used to reproduce all the computations and statistical analyses at the push of a button. This is very useful for checking and replicating results, not just for other researchers but also for the original researcher.

6.03 Practice, Ethics & Integrity: Unethical studies

Research ethics do not just concern research integrity; they also concern ethics towards participants. To understand the current thinking on ethics it's important to be aware of some of the studies that led to the formation of the ethical guidelines that we use today.

In the first half of the twentieth century several highly unethical studies were performed. When these studies were addressed in the popular media, outrage about the unethical nature of these studies led to the formation of a special committee.
The committee's report, referred to as the Belmont report, formed the basis for our current ethical guidelines. I will discuss the study that gave direct rise to the Belmont report, referred to as the **Tuskegee syphilis study** and I'll discuss a related study, the **Guatemalan syphilis study**.

The aim of the ***Tuskegee study*** was to assess the natural progression - without treatment - of syphilis, a sexually transmitted infectious disease. Participants were six hundred black, African-American men from Tuskegee Alabama. Four hundred of the men had syphilis, two hundred did not. Participants were offered health care for minor illnesses, free meals on examination days and free burials if they consented to being autopsied.

Participants were given "medical treatment" for "bad blood", a local term for different illnesses. The treatments weren't treatments at all, they actually consisted of spinal taps performed without anesthetics. The four hundred syphilis patients were never told that they had syphilis.

The study started in nineteen thirty-two, when a treatment for syphilis was still being sought. The study consisted of periodic measurements performed over a period of forty years. Half way through the nineteen-forties, around the end of the second world war, it had become clear that penicillin was an effective treatment for syphilis. And at the start of the nineteen-fifties, penicillin was widely accepted as the preferred treatment.
But until the end of the study in 1972 none of the participants were informed of this fact and no actual treatment was ever provided. Participants that were diagnosed with syphilis elsewhere were actively denied treatment or lied to and treated with placebos. A large number of participants directly or indirectly died from syphilis. Many infected their spouses; several children were contaminated in the womb and were born with the disease.

Now the Tuskegee study was not secret or performed illicitly. The study was performed by the U.S. Public Health Service. It involved many different researchers who periodically reported procedures and results in scientific medical journals. Of course it is obvious that the Tuskegee study is extremely unethical. Participants were lied to and seriously and unnecessarily harmed. They were a vulnerable group, consisting of poor, often illiterate people facing difficult economic times. These vulnerabilities were exploited in a horrific fashion.

But if you think this is the most horrendous study ever performed in peacetime you would be wrong. Just after the Second World War, the U.S. public health service performed a study in ***Guatemala***. With the cooperation of Guatemalan health services officials they actively infected people with syphilis and gonorrhea without their consent and without treatment.
Subjects were prisoners, soldiers and mental patients, coerced into participation. It seems researchers were fully aware of the unethical nature of this experiment. They knew the study would not be accepted in the U.S. Also, at the time of the study the trials at Nuremberg were taking place. These trials concerned the experiments performed by the Nazis in the concentration camps and could not have escaped the researchers' attention.

6.04 Practice, Ethics & Integrity: Research ethics

The Tuskegee study I discussed earlier, led to the formation of ethical guidelines for reviewing research proposals. Institutional Review Boards or IRB's now assess and approve research proposals involving human participants. Three ethical principles are generally distinguished: **Respect**, **beneficence** and **justice**.

Respect refers to respect for the participant's ***autonomy***. The decision to participate in research should always be made by participants themselves and should be voluntary. **Voluntary consent** can be contrasted with **coercion**. Coercion can be subtle. It can consist of offering an extremely large financial reward for participation. If the financial gain is large, then for people who have very little money, it becomes almost impossible *not* to participate. The same applies if the benefits consist, for example, of access to an experimental medical treatment that offers hope to terminal cancer patients.
A very specific form of coercion happens in most universities that offer psychology programs. In many cases first year psychology students are required to participate in a certain number of experiments for course credit. This is presented as part of their training. Alternatives to participation are offered to students, but these generally consist of very unattractive writing assignments.

Ok, back to voluntary consent: A decision to voluntarily participate can only be made if all relevant information is available to the participant. A participant should not only give consent, this consent should be ***well-informed***. An informed consent form should always be provided and signed beforehand, informing participants about the nature of the study. Of course revealing the purpose of the study conflicts with the finding that participants can react differently to the experiment if they are aware of the purpose of the study. Often, some form of **deception** is necessary to control for reactivity and demand characteristics. A review board decides whether this deception is necessary and does not cross ethical boundaries.

There are different forms of deception. Deception can occur by ***omission***: The goal of the study is not stated or formulated in very general, vague terms. Deception can also be ***active***: a cover story is provided that is entirely different from the actual purpose of the study. Or participants are given ***false feedback***. For example, participants are provided with a bogus intelligence test and they are told that they scored extremely low. The purpose could be to temporarily lower the participants self-esteem: a manipulation to see how lowered self-esteem affects people's ability, for example, to negotiate a salary.
A dangerous consequence of providing such false feedback is what's known as a ***perseverance effect***. This means participants are still affected by the deception, even after they are debriefed and the deception is revealed and explained to them. This can happen because participants might believe the researcher is just lying about the deception to make them feel better about their low scores.

If deception is deemed necessary and not harmful, then a review board might approve an informed consent form using a cover story, combined with an extensive debriefing afterwards. In all studies, participants should be made aware that they can withdraw their consent at any time during, or right after a study, and ask for their data to be removed.

Ok, the second ethical principle is ***beneficence***. Beneficence means that ***participants should not be harmed***. This principle is not as simple as it sounds. Sometimes participation carries a risk of doing harm but also a potential for doing good. The cost should always be weighed against

potential benefits. This applies at the individual level, for example when a patient participates in a study on a new cancer treatment that is a potential cure, but also has severe side effects.
But the cost-benefit analysis also applies on a broader level, for example for all cancer patients or even society as a whole. The missed benefits of not doing a study, not learning about a new cure for cancer or the cause for a societal problem, should also be weighed.

A type of harm that is perhaps less obvious is the invasion of a participant's ***privacy***. Participants should know what their data will be used for and who will have access to them. Anonymity of data should only be promised if identifying information is deleted and not even the researcher can retrace the data back to the participant.
Otherwise a researcher should be clear about the confidentiality of the data: who will have access and what will it be used for? Issues concerning confidentiality and use of data are becoming more important as more and more information of our behavior is recorded automatically.

Finally, the third principle of **justice** means that the costs and benefits of research should be divided reasonably, fairly and equally over potential participants. Specific groups should not be given preferential treatment. And reversely, vulnerable groups should not be exploited, as was the case in the Tuskegee study.

6.05 Practice, Ethics & Integrity: Research integrity

Research integrity can become compromised due to several reasons. The most serious violations of research integrity are **fraud**, **plagiarism**, **conflicts of interest** and undue influence of the researcher's **personal values**. All these threats to research integrity concern an abandonment of the basic scientific principles of openness, transparency and critical and systematic empirical testing.

Fraud refers to cases where data were either fabricated or falsified to provide false support for the researchers' hypothesis. Fabrication means the data were never collected, but were made up. Falsification means existing data were illegitimately altered. When data are fabricated or falsified to support a researcher's claims, science is damaged in several ways.

First of all, the scientific literature is contaminated with false empirical data, this alone holds back scientific progress. Researcher might pursue a line of research that *seems* promising but is in fact baseless. Honest researchers see their projects fail to replicate promising results. They might incorrectly conclude this is due to their personal lack of research skills. The project might be abandoned without the failed replications ever coming to light.

Secondly, precious funding resources that could have been spent on valid avenues of research are spent on fraudulent research and related projects that build on it, but are in fact less promising than they seem.
Thirdly, once fraud is exposed, the reputation and credibility of the field, including the majority of researchers who do have integrity, is severely damaged. This can result in more difficulty to obtain funding in a general area where fraud has been exposed. This puts honest researchers who have the bad luck to be in a related field at a disadvantage.

Fraud cases are invariably accompanied by reluctance or unwillingness to share data and research information with others. Unless researchers are challenged or even required to be open and transparent, fraud can be tempting in an academic climate were publishing positive, confirmatory results is held in high regard.
Preregistration of the research hypothesis and design, and documentation of materials, data and data manipulation form a good way to discourage fraud. If researchers know their procedures, data and analyses can be checked at any time, the risk of fraud being exposed is much more evident. Unfortunately guidelines on how to document and preregister research proposals are not always implemented and if they are these guidelines vary greatly between, even within scientific fields.

Let's move on to ***Plagiarism.*** Plagiarism is a different violation of research integrity. Plagiarism means that a substantial scientific contribution is presented as one's own by copying original text, concepts or data of others without referring to the original source.

Besides the obvious infringement on someone else's intellectual property, plagiarism contaminates the scientific literature with redundant information. If a study is plagiarized and presented as a separate, independent study, this could create the impression that a finding is more robust than it really is. Unknowing readers might interpret the plagiarized study as a successful replication, when the two studies are in fact one and the same.
Of course plagiarism often takes the form of copying small elements of other people's work, not entire studies. This is still a problem because it prevents people from having access to relevant related information in the source document, because this document is not referred to.

The large pressure on researchers might be to blame for a relatively new type of plagiarism called ***self-plagiarism***. This might seem like a contradictio in terminis, how can you plagiarize yourself? Well presenting a substantial scientific contribution that was already published elsewhere as an original contribution contaminates the literature with redundant information and makes it harder to gather relevant information if the original source is not referred to. Also when the original contribution was made with the help of co-authors, self-plagiarism means that these co-authors are not credited for their original work.

A ***conflict of interest*** is a violation of research integrity that is most frequent in the medical sciences. Researchers are funded for example by pharmaceutical companies that have a huge interest in showing a drug is effective. Whether

consciously or unconsciously, researchers can be swayed to present results more favorably. This is in part because it is also in their best interest to show positive results. Conflicts of interests cannot always be avoided, but they should at least be explicitly stated in a publication, so readers can judge for themselves what the credibility of a study is.

A final outright violation of research integrity is formed by **undue influence of personal values**. Strong conviction or personal values can blind researcher to their data and valid critiques. If researchers do not adhere to the principle of objectivity and are unwilling to accept critique or discuss plausible counterarguments based on logic and empirical evidence, then the researcher places his research outside the realm if science.

6.06 Practice, Ethics & Integrity: Questionable research practices

Fabrication and falsification of data, plagiarism and unreported conflicts of interest are considered outright violations of research integrity. But there is also a grey area of practices referred to as questionable research practices or QRP's.
Questionable research practices refer to practices that are acceptable if they are implemented objectively and responsibly, but they can abused to obtain more favorable results. These practices generally refer to selective manipulation or massaging of data and selective reporting of results.Different types of QRP's can be distinguished. I'll discuss **harking**, **p-hacking**, **cherry-picking** and **selective omission** here.

Harking is short for 'hypothesizing after results are known'. This means the hypothesis is adapted to fit the observed data. Of course researchers are allowed to formulate new hypotheses based on the data they collected. This is basically what droves scientific progress forward.
Harking becomes a *questionable* research practice if the adapted hypothesis is presented as the original hypothesis without referring to the true original hypothesis. This is a highly questionable thing to do, because results that independently confirm an a priori hypothesis can be considered relatively strong form of support for a hypothesis. But here the hypothesis was a posteriori, formed after the fact.
And hindsight is 20-20, meaning that it is easy to find an explanation that fits a specific result. Prediction is much harder. Also, adaptation of the hypothesis based on the results means that the original hypothesis was not confirmed. This failure to support a hypothesis forms useful information for other researchers who are investigating the same phenomenon. This information is lost if the original hypothesis is never reported.

Let's turn to the questionable research practice of ***p-hacking*.** A statistical test is often used to determine whether an effect - a difference between groups or a correlation between variables - is large enough to be considered a confirmation of the hypothesis.
In most cases a probability called a p-value is used to decide this issue, hence the term p-hacking. **P-hacking** refers to data manipulation or selection that makes the results - in effect the p-value - more favorable. Data manipulation could

consist of transforming scores. Data selection could consist of selecting only certain items in a questionnaire, using only certain variables or removing one of several experimental conditions.

These selection and manipulation methods can be harmless if they are performed for good reasons, for example because scores are heavily skewed, questionnaire items are unreliable or certain variables or conditions show too much missing values to provide valid results. However, sometimes these methods are employed just because it produces a more favorable p-value. The confirmation or rejection of the hypothesis thereby depends on arbitrary choices of data selection and manipulation.
The golden rule of p-hacking is that as long as data selection and manipulation are reported and arguments are provided, the reader can judge for himself whether these choices are justified. P-hacking becomes a serious, questionable problem when the data 'massaging' is not or incompletely reported.

A special form of p-hacking is **cherry picking**: reporting only results that are favorable and significant, for example only one out of three experimental conditions and only one of two dependent variables. The opposite of cherry picking is **selective omission**. Selective omission refers to the omission of non-significant results but also omission of results that contradict the hypothesis.

A last specific type of p-hacking I want to mention is **data snooping**. Data snooping refers to the collection of data exactly until results show a favorable p-value. This practice is problematic because the choice to stop is arbitrary. Suppose the results are significant - the p-value is small enough - after collecting data from seventy-nine participants. It is entirely possible that the results will be unfavorable if the data for two more participants are included. Confirmation could be based on a fluke or extreme data from one participant.
The confirmation of a hypothesis should not depend on inclusion of data from an arbitrary participant that happens to pull the results far enough in the hypothesized direction. Sample size should be determined beforehand, based on non-arbitrary estimates of the expected effect size of a treatment and the required confidence level.

Again the golden rule in all these cases is that as long as choices are reported, they can be discussed and their influence on the results can be evaluated. If choices are not reported these practices can result in serious misrepresentation and misinterpretation of results, with no way to correct these errors.

6.07 Practice, Ethics & Integrity: Peer review process

Results of empirical studies are generally disseminated, or shared through specialized scientific journals. These journals publish articles that relate to a specific research field. Traditionally these journals charge universities and others high subscription fees. However, researchers are not paid for the articles that are published and cheaper, online publishing is replacing printing of journals.

This has led people to ask why the subscription fees need to be so high. Universities pay for the subscriptions to journals that contain articles that the universities pay their employees to write. According to many, scientific knowledge should be available freely to all, without making commercial publishing companies a lot of money. In the relatively new **open access** approach, researchers, or actually the universities, are asked for an author fee if a manuscript is accepted for publication. This allows open access journals to provide the content freely to all.

In either case, closed or open access, journals need to assess the quality of a submitted manuscript and decide whether it is good enough to publish. Submitted manuscripts in reputable journals are therefor subjected to **peer-review**. This means and editor asks two to four experts in the field to review the manuscripts. In most cases the author and the reviewers remain anonymous.
Reviewers do their review work for free. Reviewing is considered a professional responsibility. For some high-profile journals the review process takes several weeks. But in most cases the process takes several months, sometimes even a year or more.

Reviewers can reject a manuscript outright, which happens often, or they can accept it 'as is', which almost never happens. Most of the time reviewers will consider acceptance if the manuscript is revised. Revisions can entail rewriting the manuscript, but also extending or redoing a part of the study. The review and revision process can be repeated one, two or sometimes even three times before a manuscript is definitively accepted or rejected.

Journals obviously want to publish high quality research articles. But how do you determine whether reviewers did a good job and accepted good manuscripts? Journals determine their quality according to the number of times their articles are cited by other articles. These citation scores are used to determine the **impact factor** of a journal. Of course publishing in a high impact journal will increase a researcher's chances of being **cited**, because people expect high quality research in a high impact journal.

For researchers it is important to publish high quality articles, measured using the number of times their articles are cited. Of course it helps to publish more papers, because this increases the chance of being cited. Unfortunately the focus on measurable research output and the increasing competition for funding and faculty positions has led to an enormous **pressure to publish**, where more is better. This trend is often described using the phrase ‘publish or perish’.

It nicely illustrates the importance of publishing, especially for researchers who are starting out and are appointed temporary faculty positions until they have proven themselves extremely successful both in publishing their research results and obtaining grants to fund their research.

6.08 Practice, Ethics & Integrity: Dissemination problems

The number of publications, citations and the impact factor determine the reputation and success of both researchers and journals. This system has the unfortunate side effect of favoring the publication of new and exciting – significant – results. This is referred to as **publication bias**, a preference to publish confirmatory, positive results.

Replications, especially non-significant ones, are generally considered less interesting and have a greater chance of being rejected. This creates a problem: Non-significant results are underrepresented in the scientific literature. This distorts our understanding of the world around us.

This is made even worse by the **file drawer problem**. Writing a manuscript and submitting it for publication is an effortful time-consuming process. Researchers often feel it is not worth the time and risk to write up and submit a manuscript that describes a non-significant result, because the time and cost involved will generally result in rejection, or at best publication in a low impact factor journal. It is more rewarding to do a new study and hope for positive results and leave the null result study *in a file drawer*.

These negative side-effects of the current system for review and publication can resolved by reforming the process on several points. A first improvement is to require preregistration of the research question, design and statistical plan. This eliminates harking, cherry-picking and selective omission.
A second improvement can be made by basing the decision to accept or reject a manuscript on preregistered research proposals, instead of completed studies. This will eliminate publication bias and the file drawer problem, even if a study produces non significant results, it is accepted and published based on the hypothesis and quality of the research design, not the outcome.
Finally, if the data and statistical analysis are made publicly available this will reduce p-hacking and hopefully reduce fraud.

Not everyone is willing to change and become more open and transparent and freely share data and materials that cost a lot of time and effort to collect. But slow steps are being taken to reform the current review and publication process and to make the dissemination of research results more open, and transparent and more publicly available.