Quantitative Methods
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.