# Quantitative Methods

## 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.