Chapter 5

Measurement design

5.1 Introduction

In the literature, a research plan or research design is generally described as the plan that details the steps that must be taken to be able to answer the research questions. A research plan is therefore a kind of 'road map', a protocol that describes the various steps to be set for conducting the research. The plan cannot foresee what the findings will be, but it can outline that a pilot study (see section 2.5) should be conducted and that based on its findings, the questionnaire and strategy to approach the sample members may be adapted. It will also specify whether it is necessary to have the plan itself reviewed by an *ethics committee* (see section 2.8) and will reserve time for approval by that committee. A research plan outlines what sample shall be drawn, what data collection strategy or instruments employed, what 'design' will be used (for instance, whether a control group will be studied, whether ethnographic data collection methods will be employed, whether respondents are to be studied repeatedly, etc.). A research plan generally also outlines the analysis methods: even though the data have not been collected yet, because of the research plan we will know what the data structure will be (we know the approximate sample size, we know the types of variables, we know whether respondents were interviewed repeatedly or not). This means that the research plan can – and should – also contain an analysis plan. Finally, it should contain a outline over the months that the study will run, showing the various activities (such as 'getting IRB approval', 'drawing sample', 'interviews', 'analysis', etc.) and a budget.

The part of research design that this chapter will be dealing with is the 'design'-part of the research plan, the research design 'in a narrow sense'. This part of research design is also often referred to as the 'measurement design', which is how we will in this book from now on refer to it, to distinguish it from the research design in a broader sense. The measurement design lays out the structure within which measurements are going to be evaluated. This will likely sound pretty vague, so let us give a few examples.

Suppose we want to find out whether the number of prosecuted cases of rape has increased since a law change in the Netherlands in 1991. In that year, the criminalized

behaviour in the legal definition of rape was broadened. The relevant sexual behaviour was extended from 'sexual intercourse outside marriage' to 'the sexual penetration of the body'. The new law thus clearly entails a much wider range of sexual acts, so it might be expected that the number of police-recorded cases of rape would go up accordingly: rape before 1991 did not legally exist within marriage, men could not be raped, and rape could be carried out only with a male sexual organ. Now, to answer our research question we take the numbers of rape cases recorded by the police before and after the law change in 1991. And it might be the case that we indeed see that the numbers increased from 1991: let us say that 3,000 cases were registered in 1990 and 3,150 in 1991. Our problem now is that while we have ascertained that an increase occurred, we do not know whether that increase occurred *because* of the law change. It could be that rape was on the rise anyway from a number of years back, and that the increase from 1990 to 1991 is not a deviation from this pre-existing trend. Or the increase could be part of natural fluctuation. Phrased differently, we observe a rise, but it is difficult to attribute the rise to the law change. Our design (where we observe before and after the law change) does not allow us to make such causal statements.

Let us discuss a second research question. Suppose we want to know whether physical maltreatment of children by their parents increases the risk that these children themselves will later become violent as well. Suppose that we have designed our study such that we investigate a group of adults who were maltreated when they were children, and a group of adults who were not maltreated as children. Now we have what is called a *comparison group* (the non-maltreated). Having a comparison group is a powerful feature of this design. It enables us to see whether we find more violence in the group of adults who were maltreated as children than in the group who were not exposed to such maltreatment. (Note that this is a measurement design we could not possibly have chosen for the first example: the law change took place for the entire Netherlands, and there is therefore no comparison group.)

Suppose now, for the child physical maltreatment example, that we find that the adults who were maltreated as children are indeed more violent: they have for instance been more often convicted for violent offences (leaving aside whether this is a valid measure of violent behaviour). Do we now know they are more violent *because* they were physically abused as children? Even though we have a comparison group that was not abused and our evidence is therefore clearly stronger than in the first example, still there are alternative explanations for the finding. It may that the abused children inherited their violent temperament from their abusing parents. If that were the case, then it would not make a difference whether children were actually abused or not as the temperament is passed on from generation to generation anyway. Or it could be that the children learnt a certain behavioural repertoire from their parents, by witnessing their parents solving conflicts or venting their anger through physical violence.

Clearly, several so-called *confounding variables* or *confounders* – variables that covary with both the childhood violent victimization and the later violent temperament – also explain the association we find between having been physically abused and later violence. All in all, using this design we cannot determine whether it is experiencing physical abuse to which later violent behaviour can be attributed.

Suppose as a last example that we want to know whether an intervention reduces anti-social behaviour. Now we can build an even stronger design. We can sample a

group of anti-social persons, divide them in two, and give (administer) the intervention to one group, and some other likely non-effective intervention, such as watching a soap series, to the other group. How do we compose those two groups? If we let our anti-social respondents choose which group to belong to, we are actually opening the door to confounding. An obvious confounder is namely 'motivation to change': the respondents who are eager to change their anti-social behaviour will likely choose to belong to the group that receives the intervention, and the unmotivated respondents will watch the soap series. If the group with the intervention turns out to be less anti-social after completion of the intervention, then this could be due to the intervention itself. But it could also be due to their different mindset, their higher motivation to change.

What we should do to prevent this – and this 'trick' is likely known to most readers – is let *chance* determine who ends up in which group. If we let chance determine who receives the intervention and who does not, it is not possible that there would be a third factor, a confounding variable, that could also explain the association between the intervention and anti-social behaviour. As only chance played a role in deciding who will undergo the intervention, there is simply no property that can be identified that is associated with receiving the intervention. Therefore, any difference between the group that underwent the intervention and the group that did not can only be explained by the intervention itself – and chance.

These three examples represent a spectrum along which designs can be ranked. The spectrum represents the strength with which statements can be made about causality, or the confidence with which we can attribute the outcome (the rise in rape cases, the prevalence of a violent temperament, a change in anti-social behaviour) to the factor we are studying. This spectrum is discussed in more detail below.

Many of the designs we discuss here are designs for quantitative studies, predominantly studies for assessing causality. Thus, in many of the examples we discuss we will speak about the effect of a certain factor X on an outcome Y. The designs we discuss will be assessed particularly for the strength with which they are able to make causal statements regarding the effect of X on Y. Design issues play less of a role in qualitative studies, as causal inferences are generally not aimed for in the same manner as they are in quantitative studies. In qualitative research, causal inferences are generally derived from context and meaning, and are generally less unifactorial. In quantitative research, we often conduct studies where we are interested in the effect, in isolation, of one or more factors X on one outcome variable Y. Qualitative studies would be employed where we want to unravel more complex multifactorial mechanisms, which cannot be reduced to a bare $X \rightarrow Y$ scheme, such as shown in Figure 5.1.

While we discussed the selection of research units in chapter 3, the choice of measurement instrument in chapter 4, and will discuss the choice of the measurement design in this chapter, it does not follow that decisions about the research plan are or should always be made in this order. In fact choices of sample, measurement instruments and measurement design are often made in an iterative process. This is because the possibilities for sampling influence the design choices, and the design that is desired given the research question also in a sense narrows the options for or dictates that certain instruments should be used.

Suppose, as an example, that we want to investigate whether and how mediation helps divorcing parents to conclude the divorce process swiftly and without conflict.

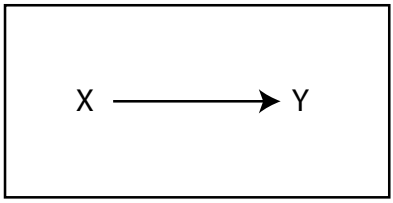
A quantitative design to investigate causality is then likely not called for. Practically speaking, we might be able to find data on a number of couples who enlisted the help of a mediator. We would then want to talk with the former partners to reconstruct what was discussed in various mediation sessions, how any potential conflicts were de-escalated, through what steps mediation was able to steer the process and let the divorcing partners construct their divorce settlement. We might learn how mediation only helps in certain settings, with certain partners. We might also uncover that in some divorce constellations mediation in fact produces worse outcomes than settlement through a family judge. The picture that we would be able to sketch would be complex, layered and contextualized. A quantitative design of the type that we will be discussing in this chapter would – practically speaking – never be able to produce such findings: the sample size is too small, and any effects depend on the type of conflict, whether custody of underage children is an issue, etc., something for which our quantitative designs are too simplistic.

In addition to such issues, time and money greatly constrain the types of design that can be chosen. If time and/or money are short, we may not be able to survey as many people as we would like, or we might not have the time or resources to develop and pilot a tailored data collection instrument. Also, it must be conceded, researcher preferences and skills play a role as well. Thus, the order in which design issues are decided upon could very well differ from the sequence in which they are presented here, and the outcome may be decided by practical constraints as much as methodological desiderata.

All the examples we have discussed so far describe essentially evaluations or assessments that are carried out *after* an intervention has taken place. Such evaluations are for that reason regularly referred to as *ex-post evaluations*. Naturally, for policy makers and drafters of laws, it is not attractive to wait until after the implementation of laws, measures or policies to hear whether they produced the expected outcome. It is good practice to attempt to estimate before implementation has even started what the likely effect of the interventions will be. Such as evaluation is referred to as an *ex-ante evaluation*. Obviously, it is impossible to assess the impact of an intervention before it has taken place, so such ex-ante evaluations rely heavily on earlier evaluations, literature review, and may employ with scenarios or simulations.

Suppose that the Council for the Judiciary has decided that in all so-called 'unilateral' divorce applications (in which only one partner files for divorce, which may be taken as an indication of conflict), a specially trained family judge will be appointed to prevent the case from escalating and to make for a smooth conclusion of the case. What impact would that have? Would the costs outweigh the benefits? Researchers could perform an ex-ante evaluation in various ways. They could conduct a literature study, investigating the impact of similar measures applied in different jurisdictions. They could interview staff at courts who already have some experience with the impact of special judges for such potentially conflictuous divorce cases. That staff might tell them that in the cases that they witnessed over the past X years, time until divorce settlement was cut by half. And the Council of the Judiciary might have statistics on the number of divorce cases that take an excessively long period of time to conclude. The researchers could then calculate the additional costs of employing such a highly trained judge, the saved costs because cases are decided upon much faster, and multiply the

Figure 5.1: Representation of relation between independent and dependent variable



difference between the two by the number of likely cases in which these special judges' skills will be employed.

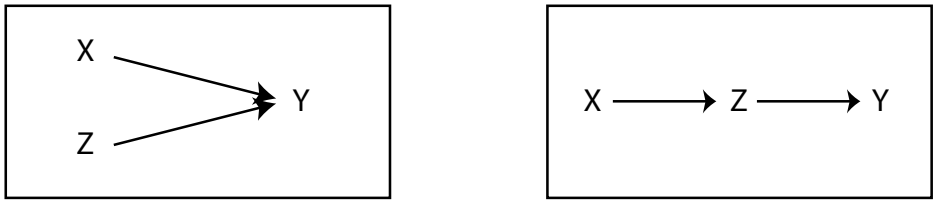
Note that such a study – while grounded in empirical observations – is still built on assumptions. It may be that once the measure is implemented, fewer conflictuous divorce cases are filed. Then, too many judges would have been trained. It might also be that more conflictuous divorce cases are filed, resulting in a shortage of trained judges. Other unexpected things might happen. It is important to realize that an ex-ante evaluation is always an estimate, an expectation. The proof of the pudding is in the eating, as the saying goes, and that saying applies very much to legal and policy interventions. The number of instances where interventions had a different impact than had been reckoned with are numerous. Sometimes this was because the ex-ante evaluation had failed to take – foreseeable – factors into account, which can be labelled as a failure to design the ex-ante evaluation properly.

5.2 Dependent and independent variables

Many of the designs we discuss in this chapter are quantitative designs for causal inference. Using the cleverest possible design, we want to make as strong as possible statements about causality. This means that in our datasets we will have variables that we think of as 'causes' and variables that are considered 'effects'. For the purposes of this book, we will defy epistemological critique and equate variables that we consider causes with the term *independent variable*. The variables that we consider effects we denote as *dependent variables*. We represent the relations between these variables as in Figure 5.1. It is a convention that dependent variables are denoted as Y. Independent variables are often written as X (which we will do too in this book). When there are several independent variables, they are denoted as X_1, X_2, X_3 ....

Obviously, dependent variables can be influenced by more than one independent variable, and independent variables (in the sense that they influence the Y variable) can themselves be influenced by other independent variables again. See Figure 5.2, where on the right-hand side the variable Z is both a dependent (predicted by X) as well as an independent variable (predicting Y).

Figure 5.2: Representation of more complex relations between independent variables and dependent variable



5.3 Causality

As said, in this chapter we will be discussing measurement designs, focusing specifically on the strength with which they allow us to make statements about causality. Before doing so, we will give a brief exposé of causality.

For X to be the cause of Y, three conditions must be met:

(1) *X precedes Y*

(2) *X correlates with Y*

(3) *there is no alternative explanation for the correlation between X and Y*

Obviously, condition 1 is logical, as we generally do not believe that causes occur after their effects. Condition 2 implies that we require that X and Y are empirically associated before we accept X as a cause of Y. Condition number 3 is crucial, as it requires that any empirical association cannot be explained by a third factor.

What is meant by this is the following. For instance, we may perceive that the number of firefighters at a fire is strongly associated with the ensuing damage. But do the firefighters *cause* that damage? The firefighters come before the eventual damage, the two are correlated, so the first two conditions for causality are met. However, and obviously, that damage is not caused by the firefighters, but both (firefighters and damage later on) are determined by the magnitude of the fire: when a large fire has erupted, many firefighters are dispatched and the damage will likely be large. We call this a situation of *spurious association* or spurious correlation: the number of firefighters and the damage are associated – not because they are themselves causally connected, but because they are each causally connected to the same, underlying *third factor*.

In this example, the magnitude of the fire is an alternative explanation for the association between the numbers of firefighters and the damage, a third factor, also called a *confounder* that blurs our analysis of the findings. Of course we would not be easily led to believe that firefighters damage fire scenes. But what of the association between (moderate) alcohol consumption and the risk of breaking your leg during skiing? A

Figure 5.3: Spurious association between number of firefighters and damage

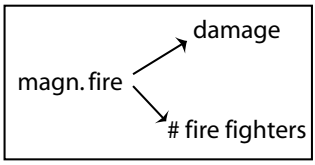
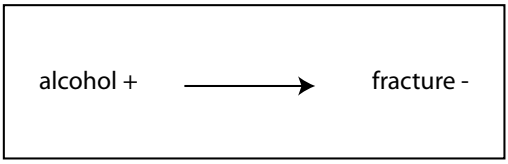


Figure 5.4: Model for the relation between alcohol consumption and risk of breaking legs



Dutch researcher a number of years ago published a study in which a clear negative association between these two variables was proven: people who drank a number of alcoholic beverages during the day, had a significantly reduced risk of fractures while skiing. The researcher said that this confirmed his theory that alcohol relaxes people's muscles, and that those who drank some while skiing would be more relaxed skiers; if they fell, they would do so in a less cramped and more supple way, thus reducing their risk of falling awkwardly and breaking something. See Figure 5.4.

After publication of the results, other researchers remarked that perhaps the association was due to the fact that people who drank more simply skied less, and thus were exposed for fewer hours per day to the risk of falling and breaking their bones. Re-analysis of the data (investigating the association between alcohol and fractures separately for those who skied 1 hour, 2 hours, 3 hours, etc. per day) proved that this was indeed the case. See Figure 5.5.

Such painful mistakes are every researcher's nightmare. Hard thinking and the exploration of all thinkable alternative scenarios that may explain the association hopefully prevent the most blatant mistakes. However, clearly we can never be sure that we have ruled out all the gazillion theoretically possible alternative explanations for the association between X and Y (condition 3). Therefore, whenever we observe an association between X and Y and even after we have ruled out all alternative explanations we could think of, still we cannot be completely sure given a correlation and the right time order that the association between X and Y is causal. Formulated differently, correlation is a necessary but not a sufficient condition for causality.

Figure 5.5: True relation between alcohol consumption and fracture risk, including exposure time

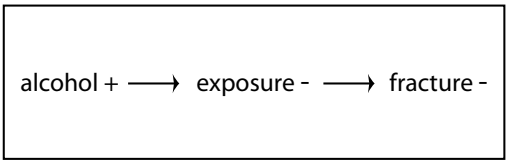
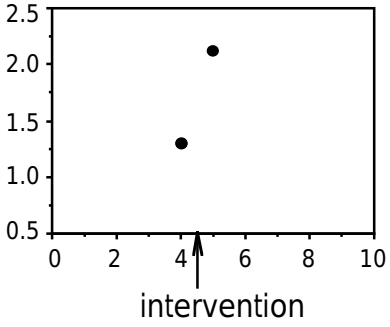


Figure 5.6: Well-being scores before and after therapy 'intervention'

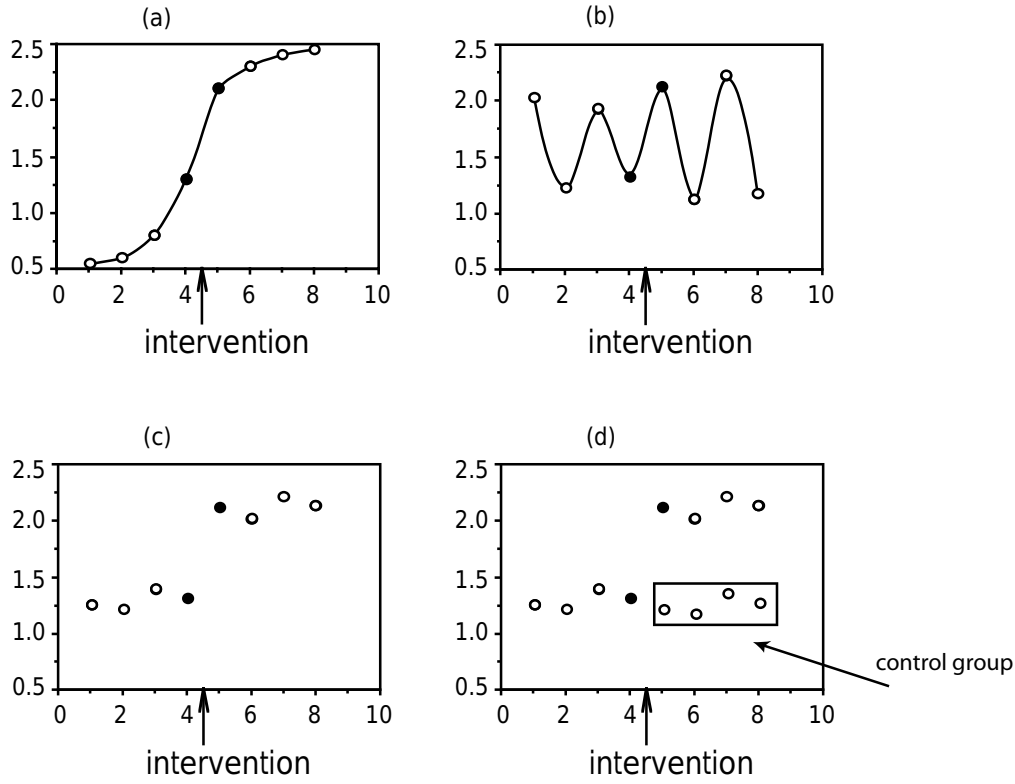


5.4 Interventions and change

Suppose that we believe that therapy increases people's well-being. We have a variable X – reflecting whether or not someone had therapy – and a variable Y – measuring well-being. If therapy works as we believe it does, then people would score lower on well-being before therapy than after. See Figure 5.6, in which we sketch a hypothetical situation where the intervention (therapy X) was administered between month 4 and 5, and where it can be seen clearly that the scores for well-being (Y) are higher after therapy than before. This is consistent with our hypothesis. We might be tempted to conclude that therapy is effective.

As argued above, we know that we cannot be sure that this improvement is *attributable* to the therapy. The increase in well-being could simply be part of a general trend: see Figure 5.7(a) where it can be seen that a steady growth in well-being set in well before therapy was administered, and that this smooth increase occurred seemingly unperturbed (perhaps there is a small additional lapse from time point 4 to 5). It could also be that the change is not more than a chance fluctuation, see Figure 5.7(b).

Figure 5.7: Development of well-being scores from month 1 to 10



Here it can be seen that the change from month 4 to month 5 is within the bandwidth of normally occurring fluctuation; in all likelihood, if the therapy had been administered between month 5 and 6 we might even have witnessed a decrease in well-being.

The graph in Figure 5.7 appears to offer the most support for an effect of therapy X: the scores hover around 1.3 before the therapy is administered, and leap to a stable level of around 2.2 afterwards. The change cannot be explained by factors such as autonomous growth or random fluctuation.

We are however still unsure, also for the situation in Figure 5.7(c), whether the change is attributable to the therapy. Suppose that we have our data from the Netherlands and that months 1, 2, 3 and 4 were very rainy, drizzly, dreary months, and that from month 5 the weather cleared and summer set in, so everyone's well-being shot up? If we wanted to know whether the increased well-being is due to the therapy or due to other factors, ideally we would like to have comparison data from a group of people who are otherwise similar (a so-called *control group*) so that they were similarly

under the influence of the weather or any other factors that impacted their well-being, and who differ from the group that received therapy (the *experimental group*) in that they did *not* get therapy, and differed from the control group *only* in that respect.

Such a situation we have sketched in Figure 5.7(d). Here we see that the control group's scores (we assume that they were equal to those of the experimental group in months 1 to 4) after the therapy remain at around the same level. It thus appears as if we may be more confident here of the conclusion that therapy increases well-being: those who do undergo therapy have decidedly higher well-being scores afterwards than those who don't. For that conclusion to be defensible, however, what we would need to be sure of is that the scores of the control group may be regarded as the scores that the experimental group would have had, had they not undergone the therapy. If that is so, the control group is then referred to as the *counterfactual* for the experimental group: the control group is the-experimental-group-without-the-intervention. However, in the situation we just described, we are not sure of that. It may be that the experimental group is composed of self-selectors, people who wanted to change, to improve their lives and who therefore elected to undergo therapy. In that case, the difference between the experimental and the control group is likely due (we do not know to what extent, but at least partially) to this difference in motivation.

These stylized examples each illustrate alternative explanations for the beneficial effect of therapy that we appeared to be seeing. Each time the issue was not whether where was an effect, but whether there was not an alternative explanation for the occurrence of the effect, other than the therapy. We went over several other explanations, such as an autonomous 'trend', and an effect of the changing of the seasons. Both these explanations jeopardized the *internal validity* of our conclusion about the effect of therapy on well-being (we discussed internal validity previously in section 2.6.2). We showed towards the end that we can only be confident that the effect is attributable to the therapy if an otherwise comparable group that did not undergo therapy performs markedly differently.

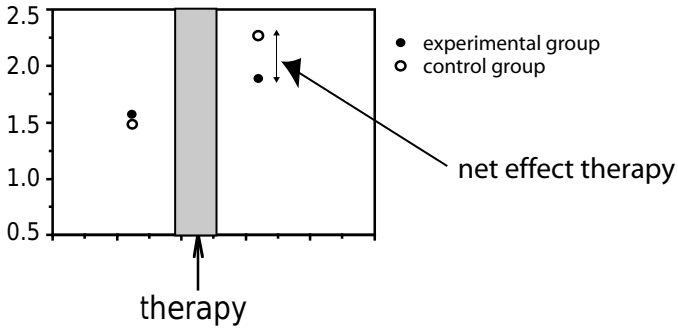
In the sections below, we will first give an outline of the most important threats to internal validity. Their discussion will be brief and summary. After that, we will rank-order designs by strength and discuss three main classes of designs, which correspond to the examples we started this chapter with. We will show that, if designs become 'stronger' in the sense that they allow for more confident statements about causality, this is because threats to validity have been ruled out.

5.4.1 Threats to internal validity

(1) *History*

Suppose I offer therapy to citizens who are afraid of armed violence. I offer my therapy to a group of 100 volunteers, all with a reasonable to serious fear of armed violence. I neatly measure their fear of armed violence before (my pre-measurement, also called *pre-test*) and after (my post-measurement, also called *post-test*) the therapy. I notice to my dismay that in the very same week in which I give my volunteers the therapy, a huge shootout occurs at a secondary school. Every citizen's fear of armed violence goes up, including that of my volunteers.

Figure 5.8: Control for threat *history* through use of control group



This external event – which influences the scores on the dependent variable – is often referred to as *history*.

It is defined as a threat to internal validity because it jeopardizes the conclusions that can be drawn about the effect of the intervention under investigation. As it is confounded with the intervention (it occurs around the same time and influences fear scores too), it affects the *internal validity* of the study: I may not conclude that the increase in fear scores (Y_2 - Y_1) is attributable to the intervention (X), as there is now a third factor (the shootout) that is an alternative explanation for the increase in fear levels.

Because of this external event, I now can judge no longer the effectiveness of my therapy. Maybe my volunteers' fear levels would have been much lower if that shootout had not occurred! If I only compare the pre- and post-therapy scores of my volunteers, there is however no way I can find that out. Obviously, I would have had no problem if I had included a comparison or *control group* in my study. The control group would then to have been comparable to the experimental group (the group undergoing the therapy). They should differ only in that these volunteers do not get the therapy. If we let chance decide who becomes a member of the experimental and control group, we are certain that there are no systematic differences between the two groups. Any differences between the two groups can then be attributed (and attributed *only*) to the therapy. There is then no alternative explanation – except chance – for differences between the two groups.

See Figure 5.8, in which the solid dots represent the scores for the treated (experimental) group and the open dots those for the control group. The figure shows that the two groups are approximately equal at pre-test and that indeed the scores for the experimental group went up, from about 1.5 to about 1.75. The scores in the control group went up much more, however, from about 1.5 to about 2.25.

Given that – if we have designed our study well and assigned respondents randomly to groups – the control group equals the experimental group without the treatment, i.e. the control group behaves how the experimental group would have behaved had they not received the intervention, we may conclude that the gain in fear generated by the treatment is -0.5 (1.75 − 2.25). In other words, we may state that the reduction in fear due to the treatment is 0.5.

The conclusion we draw is that the use of a control group eliminates the risk of history acting as a confounder. Both groups, experimental and control, are influenced by the same historical events. Comparing changes in the two, we can simply subtract any changes in the control group from the changes in the experimental group, and we will be left with the net change in the experimental group, net of any perturbances we are not interested in. Using a control group therefore means that we may – if cause X precedes effect Y and if there is a correlation between X and Y – state that no alternative explanation exists for the association between X and Y and that therefore X causes Y. The use of a(n equivalent) control group is one of the most powerful aspects of experimental designs. It means that experimental designs are strong on *internal validity*.

(2) *Maturation*

The threat we sketched previously in Figure 5.7(a) is referred to in the literature as *maturation*. Here it is not the intervention but an autonomous trend that generates the change from pre- to post-test. An often-used example is that of maths training. If I were to offer a new kind of maths training to school pupils, I would likely find that their maths skills go up from pre-test to post-test. This may be due to the maths training. But it may also be due to the fact that students mature over the course of a school year. Part of the effect may be due to the training, part may be due to that ripening or maturation.

Only by working with a control group can I determine what part of the improvement is due to the training, and what part of the improvement would have occurred anyway. Using a control group I can determine the change that the experimental group would have shown had it not received the new maths training:

$$\text{change}_{\text{control group}} = \text{autonomous trend}$$

The change in the experimental group is now due to that autonomous trend plus the effect of the maths training:

$$\text{change}_{\text{experimental group}} = \text{autonomous trend + effect maths training}$$

Comparing the two changes (for each group the difference between post-test and pre-test) gives:

$$\text{change}_{\text{experimental group}} - \text{change}_{\text{control group}} =$$

(autonomous trend + effect maths training) − autonomous trend =

effect maths training

We can write this more generically as follows. If E_1 is the average score at pre-test in the experimental group, and E_2 the average score at post-test, and if C_1 is the average score at pre-test in the control group, and C_2 the average score at post-test, then we can write the design for this study schematically as:

$$\begin{array}{ccccc} E_1 & \ldots & X & \ldots & E_2 \\ C_1 & \ldots\ldots\ldots & & & C_2 \end{array}$$

with X the intervention. This design is officially referred to as the *untreated control group design with pre-test and post-test* or as the *randomized controlled trial* or RCT, and in other literatures often as 'classical experiment'. Given that the change in the control group ($C_2 - C_1$) represents the change in the experimental group had the intervention not been administered, the change in the experimental group ($E_2 - E_1$) minus the change in the control group is the net effect of the intervention:

$$\text{net effect X} = (E_2 - E_1) - (C_2 - C_1)$$

Had we not used a control group, we would have been able to see only the gross effect of the intervention $E_2 - E_1$, being the combination of the effects of the maturation and the maths training, without being able to separate the two.

(3) *Testing*

It is known from the literature that administering tests or questionnaires by itself changes people's ideas, skills, and thus possible scores on the next administration of that same questionnaire. A relative of the author once had to be operated, and was asked, before receiving the anaesthetic, to fill in a small test. After waking up she filled in the same test. To the surprise of the researcher, she had better scores – a few hours after the anaesthetic – than before. She, being a well-trained social scientist explained to him that this was due to *testing*, the phenomenon that people's scores are affected by the completion of tests: she simply learned from filling out the test the first time.

As the attentive reader will have gathered, this threat to validity is also effectively countered by using a randomized control group. If testing affects the scores, then it will do so equally in the control and experimental group and we can again find the net effect of the intervention by comparing the changes in the control and experimental group.

(4) *Instrumentation*

By *instrumentation* is meant that apparent changes from pre- to post-test may actually be generated by changes in the measurement instrument. Thus, one observes a change in respondents' scores from pre-test to post-test, but the different

scores are not due to people changing, but due to the instrument changing from pre-test to post-test. If one could, one would likely always keep one's instrument identical across multiple waves of the same study, across respondents, etc. But sometimes this is not possible: political parties have changed from one wave to the next, the meaning of religion and church attendance has changed from 1950 to the present, and the like. Again, this threat of instrumentation is countered by using a control group. If instrumentation affects the scores, then it will do so equally in the control and experimental group and comparing the changes between the control and experimental group will again give us the net effect.

The first four threats that we discussed are therefore all effectively countered if we use a control group. In each instance, the 'distortion' occurs in both the control and experimental group and any threat is simply subtracted by computing the net effect $(E_2 - E_1) - (C_2 - C_1)$. It may be that several threats occur simultaneously, or that other disturbances also affect the data – but all of that does not bother us, as the design is such that we will always be able to distil the net effect of the intervention:

$$\text{change}_{\text{experimental group}} - \text{change}_{\text{control group}} =$$

$$(\text{history} + \text{maturation} + \text{testing} + \text{instru.} + \text{any other} + \text{effect intervention}) -$$

$$(\text{history} + \text{maturation} + \text{testing} + \text{instru.} + \text{any other}) =$$

$$\text{effect intervention}$$

For the threats to internal validity we subsequently discuss, the use of a control group as such is not always sufficient. These threats either bear on other issues that may affect the usefulness of a control group, the choice and retention of group members, or they deal with more practical issues, such as whether control and experimental group members know of each other's existence.

(5) *Selection*

We say that *selection* is a threat to internal validity when the experimental and control group have not been put together randomly. Whenever that is the case, we cannot rule out that a third factor (the factor that influenced who became a member of which group) can also (partly) explain the difference between the experimental and control group.

Selection definitely occurs if one lets sample members decide themselves whether they want to be a member of one group or the other. It may also occur if we work for instance with existing school classes. Even though it is hard to envisage how exactly conclusions would be affected, knowing that there are other systematic differences between the groups (for instance, different teachers) is sufficient to be aware that internal validity is compromised.

If the control and experimental groups have not been formed by a random process, it may happen that one or more of the threats discussed above occurs in one group but not in the other. If my fear of armed violence experimental group is based in Alkmaar, and my fear of armed violence control group in Alphen aan

den Rijn, where a massive shooting incident took place in a shopping mall in 2011 just when I was offering my therapy, it is likely that this history type of effect plays a larger role in the control group than in the experimental one. This might then even make us overestimate the effect of the therapy to combat fear of armed violence.

(6) *Attrition*

As we discussed in section 3.6, attrition is generally non-random: those who drop out of a study generally have different characteristics than those who remain. In the context of experimental design, attrition or drop-out (also called *experimental mortality*) may cause groups to become incomparable, as attrition may differ from one group to the other. This means that if we start out with two comparable groups, control and experimental, beautifully formed by chance, attrition will generally cause these groups to become non-random. Attrition is reported to occur more easily in the control group as respondents are aware that they have not been included in the group that receives the experimental intervention. Then, the same problems as outlined just above under 'Selection' emerge.

(7) *Insufficient separation between groups*

Suppose that we are conducting a study within a large firm. We are trying to train employees to work more efficiently. Now, we try to evaluate the training according to the state of the experimental art – that is, we have formed an experimental and control group from employees within one and the same firm. However, having trained the experimental respondents, there is a risk that the training in a sense 'seeps through' to the control respondents: they meet each other afterwards at the coffee machine, in the elevator, over lunch. Control employees might even witness the experimental employees employ all these handy tricks, and decide that they want to become as efficient.

We call this situation one of *insufficient separation between groups*. Because the intervention also reaches the control group, leaks to the control group members, we are likely to underestimate its effect.

A solution is hard to find in this case. Forming a control group out of employees from, say, a different branch will make the control and experimental group less comparable, and thus generate other problems.

(8) *Perceived differences between groups*

Continuing from the example we just discussed, it might also occur that a certain envy is generated by this design. Suppose that the experimental respondents are trained in a nice location and share fancy lunches together, while the control group respondents have to plod on in the same old dreary office. Examples have been reported where the control respondents gave up doing their job decently altogether (resulting in the effect of the training being overestimated), as well as the opposite where the control respondents work so hard to show off that it would have to be concluded that the training was useless (thus meaning that its effect is underestimated).

(9) *Regression towards the mean*

To give an example of the last threat that we discuss, suppose we want to investigate the effect of, again, a maths training. We have a group of children who have had consistently low scores on maths tests, and these form our experimental group. Of course, we know we must have a control group as well, so we administer a maths test to all other children in the school, and we take the children scoring lowest on that test as our control group. Regression to the mean is now likely to occur. But what do we mean by that? It is best understood as follows.

It is likely that in our lowest scorers who formed the control group, a few scored low on the test just because of some quirk, they had an off-day, a headache, they were madly in love. They were not consistently low scorers, otherwise they would have been members of the experimental group.

Now if we compute the scores of this group on the post-test, it is likely that some of those accidentally low scorers will at post-test no longer have their off-day or their headache or bout of love-sickness. They will score higher at C_2 than at C_1. Their scores have regressed ('gone back') to the group mean. This means that we – in this example – will underestimate the effect of the maths training.

Regression towards the mean regularly occurs, also outside the context of experiments. It was first described by Francis Galton, half-cousin of Charles Darwin, an extremely broad scientist and an avid traveller to some of the remotest corners of Africa.

5.5 Types of (experimental) designs

Other types of measurement designs than the ones we discussed are possible too. For example, we may interview a sample of respondents once, and ask them for their opinion about a number of topics. Such a design does not allow for causal statements. The designs we discussed so far all have in common that they are structured such that we can make statements about the causal effect of an independent variable on a dependent variable. We have a pre-test and a post-test, which means that we measure at least twice, and always have a comparison to gauge our results. Whenever we want to make causal statements about the effectiveness of an intervention, an experimental design is needed.

Experimental designs vary – as elaborated above – by the strength with which they allow to make causal statements. We saw that designs without a control group are vulnerable to threats of internal validity. Including a control group is already a huge improvement. The strongest designs are, however, those that have randomly formed experimental and control groups. In that case, it can be excluded that any systematic difference between the groups can explain differences between the groups: as the groups were formed randomly, there are no systematic differences – except the intervention. A difference in scores can therefore be tied to the intervention, and to the intervention only.

In criminology, the so-called *Maryland Scientific Methods Scale* or MSMS (Sherman et al., 1998) is often used as a reference to gauge the strength of particular designs.

It has five levels. The levels can be summarized as follows:

(1) At this level a correlation has been established between an intervention and a dependent variable at a single point in time.

(2) A temporal sequence between the intervention and the outcome has been observed, or there is a comparison group without demonstrated comparability to the treatment group.

(3) A comparison between two or more comparable units of analysis, one with and one without the intervention.

(4) Comparison between multiple units with and without the intervention, controlling for other factors, or using comparison units that evidence only minor differences.

(5) Random assignment and analysis of comparable units to intervention and comparison groups.

While the first level is fairly vaguely described, level 2 describes a design where a control group is present, albeit an incomparable one. Sherman et al. (1998) use as a rule of thumb that designs below level 3 are unfit for evaluation research. Any design that does not meet the criteria of level 3 or higher should therefore be disregarded for causal inference. Level 3 describes a situation where a reasonably comparable control group is present. Level 4 describes the situation where the experimental and control groups differ only slightly. At level 5, we have a truly randomized design, an RCT. This means that the Maryland Scientific Methods Scale rejects for causal conclusions all studies without a well-comparable control group.

In a completely randomized design, the researcher exerts control over all aspects of measurement: s/he assigns respondents to groups, (often although not always) administers the intervention, and conducts the measurements. Designs with non-randomized control groups (levels 2, 3 and 4) are in the literature referred to as *quasi-experimental designs*. In a quasi-experimental setting, there is less control: while the researcher may still administer the intervention and conduct the measurements, s/he works with existing groups, such as classrooms or companies.

Both the experimental and quasi-experimental designs look schematically as follows:

$$\begin{array}{ccccc} E_1 & \ldots & X & \ldots & E_2 \\ C_1 & \ldots & \ldots & \ldots & C_2 \end{array}$$

but differ in the nature of the control group (randomized or not). The MSMS disaggregates quasi-experimental designs into those with very well comparable control and experimental groups (level 4), those with comparable control and experimental groups (level 3), and last those with actually quite incomparable control groups (level 2).

Other types of designs than the randomized controlled trial are possible. The following is called the one-group pre-test-post-test design:

$$E_1 \quad \ldots \quad X \quad \ldots \quad E_2$$

This is the design that we depicted in Figure 5.6. It is as we showed a weak design. A better design is the so-called *time series design* that we depicted in Figure 5.7(a), as with this design we will be able to detect maturation, as well as testing and instrumentation. It does not control for history occurring simultaneously with the intervention:

$$E_1 \quad E_2 \quad E_3 \quad \ldots \quad X \quad \ldots \quad E_4 \quad E_5 \quad E_6$$

If the intervention in such an 'experimental-group-only study' is not administered by the researcher, the study is denoted as an *observational* or *passive observational study*: the researcher does not administer the intervention and does not assign respondents to groups; all s/he can do is observe the data.

Luckily, some intermediate solutions are possible that improve upon the situation of having only observational data. One such possibility is to seek, for every member of the 'experimental group', one or more observational units that are seemingly identical to the experimental group unit – but for the 'intervention'. For instance, if we were studying complicated divorce settlements, we might want to investigate whether mediation helps to conclude cases more swiftly than normal. Obviously, we are unable to randomly assign cases to mediation or not, as this is something that is up to the parties to decide. Some former couples have decided that they may seek mediation, while others have not, and we can only observe this happening. We can, however, for each couple that used mediation, try to find a similar ('matching') couple – in terms of relevant dispute and divorce characteristics – that did not use mediation. If the mediated and non-mediated are similar with regard to a host of pertinent characteristics (financial claims, presence and age of children, employment of spouses, other relevant circumstances of the case), we could compare differences between the similar cases settled with and without mediation, and see whether the mediated cases were indeed concluded faster. Such a design is called a *case-control design*. The procedure that we apply to arrive at such a design is called 'matching'.

Matching is very broadly used when it is impossible to randomize the intervention we want to assess the effect of. It definitely improves upon the strength of our conclusions, but does not eliminate confounding. For instance, if we were able to match perfectly on a certain number of background variables, we would be sure that the experimental and control group are identical on those variables, and that any differences between the two are therefore not attributable to these factors anymore. However, we are still unsure whether there are not other, unmeasured ('covert') factors that we have been unable to match on, that are confounded with the 'intervention' we are studying. This is a first disadvantage of matching. Second, it should be noted that matching does not by itself eliminate confounding on the matching factors (for more on this, see Pearce, 2016). A last and more practical disadvantage of matching is that if we want to match on a large number of characteristics (which we would always want to do because then we eliminate the risk that these factors blur our conclusion on the relation between the intervention and the independent variables), it may quickly become impossible to do so. It may simply be impossible to find a matching case that is identical as to type of conflict, age of partners, presence and age of children, socio-economic status of parties, and so forth. Innovative methods such as propensity score matching (see Bijleveld et al., 2015) provide tools to deal with this, but even so matching remains a second-best

solution that one resorts to only when one has no other option. This is unfortunately often the case, including in empirical legal research.

5.6 Longitudinal and cross-sectional designs

Whenever measurement units are observed repeatedly, we call the study *longitudinal* to denote that it stretches over time and, crucially, is designed to measure change. Longitudinal studies are often contrasted with cross-sectional studies, where measurements are made at one point in time, and the aim is not to observe or explain change.

Longitudinal designs can be classified into prospective and retrospective designs. Of the two, the prospective design is by far the strongest one. We will illustrate this with an example. Suppose that we study a sample of male child sex abusers. And suppose that we note that almost all of these sex offenders were themselves sexually abused as a child. We might then be tempted to hypothesize that child sexual abuse is a risk factor for sex offending against children. This statement is however based on *retrospective* data: because of our design, we only observe those who became a sex offender, but not those who were abused but did not become a sex offender. If we want to be sure that sexual abuse victimization is a risk factor in the aetiology of sex offending, we would have to investigate a sample of sexually abused children (as well as, preferably a sample of not sexually abused children), following them into adulthood. This is what is referred to as a *prospective* study, prospective because measurements are not obtained looking back, but looking 'forward', so to speak. The big advantage of this prospective design is that we will also be able to see the negatives: the number of children who have been abused who do not become sex offenders.

While prospective studies are methodologically much stronger than retrospective studies, they are – not surprisingly – much more costly and may take a very long time to complete. Prospective research is much more expensive as it takes so long: it is extremely costly in terms of database maintenance, follow up of sample members, and staff costs. It often takes too long for policy purposes too.

5.7 Vignette studies

A special type of RCT is the so-called *vignette study*. In a vignette study hypothetical cases (or 'vignettes') are offered to respondents. The properties of these cases have been altered purposely from case to case, with the relevant properties thus, as it is called, manipulated.

De Keijser & van Koppen (2004) used such a vignette study. They investigated whether in the sentences that judges mete out in the Netherlands these judges show evidence of so-called 'compensatory punishment'. The authors report that many lawyers are convinced that judges tend to give more lenient punishment in cases where proof is not very strong: the idea is that judges 'compensate' for the fact that the evidence is not overwhelming by giving shorter sentences. It would be very hard to investigate this assertion in real court cases as so many factors play a role in judicial decision making that could act as confounders. The authors therefore constructed three types of court

Table 5.1: Results of experiment on compensatory punishment

crime	proof	sentence in months	*t*-value
serious assault	strong proof	28.7	1.85 (ns)
	weaker proof	32.9	
burglary	strong proof	5.4	0.66 (ns)
	weaker proof	5.7	
simple assault	strong proof	3.1	1.40 (ns)
	weaker proof	2.7	

ns = not significant

cases, a simple assault, a serious assault and a burglary case. They changed the cases such, that for each type of offence there was a description where the proof was sufficient (otherwise judges would acquit) but not overwhelming, and one case where proof was very strong. In this manner, the authors *isolated* the factor they were interested in: strength of proof. Each time the case description was identical, the only aspect that varied was how strong the proof was. Changes in sentence length could therefore be attributed only to strength of proof.

In total 36% of judges that were asked to participate consented. Each judge was randomly given one case with strong evidence and one case (with another offence, otherwise judges would understand what the authors were after) with weaker evidence. The results of their study (with the properties of the fictional cases and average sentence length in months) are in Table 5.1.

The authors concluded that their experiment did not show any evidence of compensatory punishment. They also note, however, that the cases they presented judges with are 'artificial' cases and lack the depth and complexity of real court files (De Keijser & van Koppen, 2004). They therefore do not want to rule out that in reality compensatory punishment occurs (see also section 5.8 below). With these reservations, the authors concede that while the internal validity of their vignette study is high, external validity may be less.

5.8 Pros and cons of various designs

The experimental designs we discussed above, particularly the completely randomized design or RCT, are in a sense the gold standard of all evaluation research. We reiterate – possibly to the boredom of the reader – that only this design gives us hard evidence of causality. That being said, and being true, experimental designs have their drawbacks too.

First, experimental designs score high on internal validity, but their so-called *external validity* – the extent to which these findings are replicable in other settings – is disputed. A situation where a researcher divides two groups randomly in two is hardly

a 'naturalistic setting' and examples have been reported in the literature where an effect of an intervention could be proven in a lab setting – but not outside in the 'real world'. This *external validity* is therefore a special kind of generalizability – not from a sample to the population, but from one setting to another setting. This weak point of experiments is also referred to as lack of *ecological validity*. Thus, even when conclusions are as 'hard' as cast iron within the research setting, they will be of little use if in real-life applications they would not emerge. This has led some scholars to prefer more naturalistic experiments over more controlled ones.

The ecological validity of controlled experiments may be compromised by several factors. One such factor is the so-called *Hawthorne effect*. This effect is named after a factory in the US where the effect of lighting on worker performance was investigated. The (possibly apocryphal) story goes that whichever way the lighting in the factory was altered, work performance always improved. This finding was explained by supposing that performance went up because the workers knew they were being investigated.

Another such threat to external validity is the *placebo effect*, also well known from pharmaceutical research. By this is meant that just from receiving some medicine (whether it has active ingredients or not) patients report feeling better. This effect can easily be countered by offering control respondents an intervention that has all the looks of the experimental intervention, but lacks the relevant elements. Thus, if we wanted to investigate the effect of a training video on secondary school students' bullying behaviour, we would, to counter the placebo effect, have to offer the control group a video too, although on an unrelated, neutral topic.

The last threat to external validity we will discuss (a few more have been listed in the literature) is the so-called 'Biotex' effect (Bijleveld & Van der Geest, 2021). Now this may seem an odd name, as Biotex is a special kind of washing powder which Dutch housewives love because after soaking hopelessly dirty clothes in it for a while, these become sparkling clean when washed in the washing machine afterwards with regular detergent. This Biotex effect has also been observed in experiments, in the sense that an effect of the intervention can only be demonstrated after respondents have been 'sensitized' to the intervention by a pre-test. Of course, in the real world, the intervention would generally not be administered after a pre-test and then the effect would not emerge. This threat can be detected by adding a second control and experimental group to the experiment and not administering a pre-test for these additional groups.

Secondly, not all interventions can be – as it is sometimes phrased – manipulated by a researcher. It is impossible to randomly assign child maltreatment or heroin use. The impact of such events can generally only be studied non-experimentally in naturalistic, observational settings. And in this, science is sometimes aided by natural events or manmade interventions that make it possible to study the effects of the upheaval they cause. Such occurrences are called *natural experiments*. We discuss such natural experiments in somewhat more detail in the next section.

5.9 Natural experiments

Titiunik (2021) ranks natural experiments, in terms of the extent to which they can provide proof of causal effects, in between the RCT and a quasi-experimental study. In an

RCT the researcher has *assured* through some sort of randomization procedure that observation units in the experimental and control conditions do not differ systematically – in other words, that only chance differences exist. In a quasi-experimental study, the researcher has not formed the groups, a different entity has done so, and as such we cannot be sure that the groups have been randomly formed and we must therefore reckon with the possibility that confounders obscure our picture.

In a natural experiment, according to Titiunik (2021), the treatment assignment mechanism is neither designed nor implemented by the researcher, is unknown to the researcher, but is probabilistic by means of an external event or intervention. It is generally assumed that that external event leads to random assignments of persons to groups. Because the researcher does not exert control over the randomization, such assignment is referred to as 'as-if' randomization: we can only assume that assignment was indeed random. One important additional condition for such an observational study to be classified as a natural experiment is that the respondents who are 'assigned' to the experimental and control groups do not exert any control over this assignment: they also must be unaware of the mechanism, and they should be unable to anticipate the intervention.

Natural experiments are often found upon the introduction of administrative or legal changes; often the 'collateral', unintended effects of such laws are studied employing the fact that the 'intervention' may be considered to qualify as 'as-if' random.

An example from life-course criminology illustrates such a natural experiment. In the Netherlands, all men born in 1959 were exempted from military service. The Dutch government had introduced this one-year exemption as it could not handle the large numbers of men who were being drafted. If one wanted to study the impact of military service on life outcomes such as marriage chances or earnings, one could not do that validly by comparing the outcomes of men who have served and who have not: medical and psychological checks for admission into service are likely confounded with characteristics that are associated with these life outcomes as well. However, as none of the men born in 1959 were drafted, through an entirely random mechanism, and as there is no reason to suppose that the cohort born in 1959 differs systematically on any relevant characteristics from men born in 1960 or in 1958, Van de Weijer & Bijleveld (2016) could exploit this random intervention and compared the criminal records of men born in 1959 (of whom no one had been drafted) with those born in 1960 and in 1958 (of whom some were drafted). They found no difference in the percentage of men who had a criminal record for these birth years, meaning that the military service did not impact criminal careers. They did however find that men who had not been drafted were more likely than those who had been drafted to follow in their father's footsteps, that is, to also become an offender if their father was an offender. The authors interpret this as suggesting that military service may 'break' the intergenerational cycle of transmission of criminal behaviour.

A famous example also using military records is the study by Angrist (1990). Research using administrative records had shown that men who had served in the Vietnam war earned less than men who had not served in the Vietnam war. The issue at stake was whether that was so *because* they had served in the war. Again, because various characteristics of the men could co-vary with both the choice to sign up or not and with lifetime outcomes such as earnings, it is hard to attribute any lower earnings to

the military service in Vietnam. However, five draft lotteries had been held during the Vietnam war period, from 1970 up to 1975. And within these years, a number of cohorts could be identified that had groups of drafted and non-drafted men, formed randomly. Angrist showed, using Social Security administrative records, that the earnings of white veterans within these groups were approximately 15% less than the earnings of comparable nonveterans.

Natural experiments are also found when disasters occur. One such disaster was the so-called 'Hunger Winter' in the Netherlands, the 1944–1945 winter that was extremely cold and in which particularly the west of the Netherlands under Nazi occupation suffered severe food shortages. In the winter and spring of 1944, after a railway strike, the Nazi German occupiers limited rations such that people received as little as 400–800 calories per day. It was known that babies born during or right after this period had lower birth weights. Researchers studied the long-term consequences of this famine in people born or conceived during the Hunger Winter. The design appears to meet, to a large extent, the criteria for a natural experiment: the treatment assignment mechanism is neither designed nor implemented by the researchers, is an external intervention, that is unforeseeable and outside of the control of the subjects of the intervention. One could of course argue that the experimental group misses out on parents who managed to leave the west of the country (although that was very hard to do), and that children born earlier may have been exposed to the famine too. The design ranks, all in all, with its 'as-if' randomization, somewhat below an RCT in terms of strength to prove causal relations.

The researchers studied three groups of people. Firstly they studied children born during the Hunger Winter or right after (and thus conceived during that period), who could be called the 'experimental' group; these children had been found through the registries of the midwifery schools in Amsterdam and Rotterdam, and the delivery ward in the Leiden University Academic Hospital. These were therefore children born in the west of the country where the famine was worst. A second group consisted of children born in the same clinics, but in 1943 or in 1947, so before or after the Hunger Winter; this group served as a control group of children born in the same location, but not exposed to the famine. See for more on the samples Rooseboom et al. (2006).

A subsequent study showed that women exposed to the famine during mid- to late gestation had babies with significantly reduced birth weights. Babies whose mothers were exposed only during early gestation had normal birth weights. However, the latter grew up to have higher rates of obesity than those born before and after the war and higher rates than those exposed during mid- to late gestation, with the difference in weight on average almost 5 kgs. At ages 56 to 59, both men and women exposed to famine during the early stages of gestation performed worse on a selective attention task (De Rooij et al., 2010). Thus, the studies have found both physical and mental long-term health consequences.

Natural experiments have not been employed often yet in empirical legal research. One set of studies on sex discrimination cases is discussed in detail by Erikson (2022). These studies all investigated the effect of the presence of a female judge in sex discrimination cases, i.e. cases where discrimination against an individual is claimed because of gender identity, including transgender status, or because of sexual orientation. The author studied a database of 435 sex discrimination cases, containing US Appellate

Court decisions from 1995 to 2002. These are cases where a decision is arrived at by a panel of judges. And one reason why the impact of the gender of the judges can be studied using these data is that judges should be assigned *randomly* to panels: there are therefore no systematic differences between the judges that may co-vary with the types (and therefore possibly the outcomes) of sex discrimination cases.

Earlier research had found that male judges were about 10% more likely to side with a female plaintiff if they served with a female colleague on the panel, a finding that was highly statistically significant (see Erikson, 2022). The author, before embarking on his analyses, tested first whether assignment of judges to cases could actually be assumed to be random ('as-if' randomness). For that, he investigated whether male and female judges differed in terms of the earlier verdict in the case they were assigned to. He found that that was not the case, so that random formation of panels with male and female judges could be assumed.

Erikson (2022) next improved on previous research with two additional control steps. First, he compared cases with an all-male panel (about two-thirds of the panels) with cases where the panel contains a female judge, each time within the same time period. This rules out the alternative explanation that more female judges sit on panels in more recent years, and that any difference in favour of sex discrimination plaintiffs is actually a period effect (judges become more responsive to sex discrimination cases in more recent years). Erikson also compared only pairs within the same appellate circuit to rule out any confounding on this variable. Second, he compared panels (with and without a female judge) pairwise, with the judges in the pairs of control and experimental panels having the same distribution of political affiliations. This is to rule out the alternative explanation that it is not a judge's female gender which impacts judgments in sex discrimination cases, but the fact that female judges more often vote Democrat.

Erikson also exploited an attractive property of the data, namely that male judges sometimes sit on panels with only male judges and sometime on panels with a female judge as well. Erikson could therefore investigate not only effects across panels of judges, but also *within* judges over time, in longitudinal fashion. He investigated judgments of male judges before and after having served with a female judge, and found that male judges after having served on a panel with a female judge do not become more likely to vote in favour of plaintiffs in sex discrimination cases. Erikson's study (2022) confirms earlier research, but he reports a larger effect size, of around 13%. It should be noted that Boyd, Epstein, & Martin (2010) found such gendered judicial effects also (but only) for sex discrimination cases. These authors therefore conclude that in the USA the presence of women in the federal appellate judiciary rarely has an appreciable empirical effect on judicial outcomes.

5.10 Further reading

The 'classic' read on experimental designs is Cook & Campbell (1979), with Shadish, Cook, & Campbell (2002) its successor. Farrington (2003) is a lucid and much cited piece on standards for evaluation research, mostly focusing on crime and criminal justice.

Chapter questions

1. What are the necessary and sufficient ingredients for a randomized controlled trial? (section 5.4.1)
2. Argue what confounders can be detected in a design with only pre-test and post-test (section 5.4.1)
3. Why are history, maturation and testing all eliminated as confounders in a design with randomly chosen experimental and control groups? (section 5.4.1)
4. Argue why selective attrition is more likely to occur if no placebo is included in an experiment and why that constitutes a threat to internal validity (section 5.4.1)
5. Argue whether a vignette study meets all the criteria of a randomized controlled trial (section 5.8)
6. Do the same for a natural experiment (section 5.9)
7. What is meant by 'as-if' randomness? (section 5.9)
8. Argue whether the introduction of a new law constitutes a natural experiment (section 5.9)