JAN VANHOVE

QUANTITATIVE METHODOLOGY

AN INTRODUCTION

AUTUMN 2024

UNIVERSITY OF FRIBOURG

jan.vanhove@unifr.ch
https://janhove.github.io

Preface

The principal goal of this class is to impart the fundamentals of quantitative research. Our focus will be on **drawing causal conclusions** from data—when are causal conclusions licensed, what are typical pitfalls when drawing causal conclusions, and how can you design and optimise studies that avoid these pitfalls? Throughout, it is imperative that you not only understand the recommendations given in this booklet but also the **logic** behind them. There are two reasons for this.

First, you need to know when the recommendations apply and when they don't. You might get away with just memorising the recommendations and their scope for now. But in a couple of months, you're bound to apply them where they don't make any sense. Understand the logic behind the recommendations, and you'll be better able to weigh your options.

Second, many researchers in the social sciences—even seasoned ones—operate on rules of thumb, so *they* inevitably end up applying recommendations they've picked up somewhere in situations where they don't apply. You need to be able to make an informed judgement about the research carried out by others and cogently argue for this judgement. A simple *I took this class that taught me that you shouldn't control for colliders* won't do. (You'll learn about colliders and why you shouldn't control for them soon enough.)

In this class, we'll cover the contents of this lecture script. In addition, we'll spend two sessions on questionnaire design. Furthermore, in the homework assignments, you'll learn how to visualise data using state-of-the-art software. These graphing assignments are available from `https://janhove.github.io/graphs`.

I occasionally refer to some blog entries I wrote. These links are clickable in the PDF version of this booklet, but in case you're reading this from paper, all blog entries can be found at `https://janhove.github.io`.

Jan Vanhove
`jan.vanhove@unifr.ch`
`https://janhove.github.io`

Contents

1
Association and causality

1.1 Two examples

Below are two examples of empirical findings and some possible conclusions. Answer the following questions for both of these examples.

1. Do the conclusions follow logically from the findings? If not, what are some plausible alternative explanations for the findings?

2. Which additional findings would strengthen the conclusions?

3. Which additional findings would call the conclusions into question?

Example 1.1 (Receptive multilingualism in Scandinavia)**.** When talking their respective native languages, Danes understand Swedes better than the other way around. Furthermore, Danes like Swedish better than Swedes do Danish (e.g., Delsing & Lundin Åkesson, 2005).

Conclusion: Danes understand Swedes better than the other way round because they like the language better. ♢

Example 1.2 (Content and language integrated learning)**.** Pupils in *Content and Language Integrated Learning* (CLIL) programmes in Andalusia perform better on English proficiency tests than other Andalusian pupils (Lorenzo et al., 2010).

Conclusion: Taking CLIL classes improves pupils' English proficiency. ♢

In both examples, an **association** of some sort is found in the data, and a **causal explanation** of this association is put forward: Not only do Danes both understand and like Swedish better than Swedes do Danish (association), it's suggested that one reason why they understand the other language better is that they like it better (explanation). Similarly, not only do CLIL pupils in Andalusia outperform non-CLIL pupils (association), it's suggested that they outperform them *because* of the CLIL programme (explanation).

Uncovering associations and drawing causal conclusions from them is a key goal in empirical research. But it's also fraught with difficulty: after a moment's thought, you'll often be able to come up with alternative explanations for the findings. To the extent that there exist more, and more plausible, alternative explanations, the causal explanation proferred becomes more tenuous: The causal claim may still be correct, but in the presence of competing explanations, it can't be *shown* to be correct—that is, there isn't much **evidence** for the claim. A key goal when designing an empirical study is to reduce the number and the plausibility of such alternative explanations.

1.2 Definitions

Definition 1.3 (Association)**.** Two factors (or variables)[1] are **associated** if knowing the value of one factor can help you hazard a more educated guess about the value of the other factor. ♢

That's a mouthful, but convince yourself that the following are examples of associations:

- a person's size in centimetres and their size in inches;
- the time of day and the temperature outside;
- a person's height and their weight;
- a person's shoesize and the size of their vocabulary in their native language;
- a person's age and their father's age;
- a person's nationality and the colour of their eyes.

Five remarks are in order:

- Associations work in both directions: knowing the time of day allows you to venture a more educated guess about the temperature outside than not knowing it, but also vice versa.
- Associations needn't be linear (e.g., the relation between weight and height levels off after a certain weight).
- Associations needn't be monotonic, e.g., the relationship between the two variables can go up and then down again (as in the time of day/temperature example).

[1] I use these terms interchangeably. Sometimes, factors are constant rather than variable in the context of a study, but let's save our pedantic inclinations for other things.

- Associations needn't be perfect (e.g., there's a lot of variation about the general trend for taller people to be heavier).
- Associations can be found between variables that aren't typically expressed numerically (e.g., eye colour and nationality).

Typical examples of associations in research are mean differences between groups and correlations.[2]

As for **causality**, a common-sense understanding will be sufficient for our purposes. But when in doubt, you can turn to the following broad definition:

Definition 1.4 (Causality)**.**

> "We say that there is a *causal relationship* between [two variables] D and Y in a population if and only if there is at least one unit in that population for which intervening in the world to change D will change Y . . . ." (Keele et al., 2019, p. 3)

◇

Three remarks are in order:

- Saying that D causally influences Y doesn't imply that D *alone* causally influences Y. (You can get lung cancer from smoking, but also from exposure to radon, air polution or just genetic bad luck.)
- Saying that D causally influences Y doesn't mean that changing D will result in a change in Y for *all* members of the population. (Some non-smokers get lung cancer, and not all smokers get it.)
- Saying that D causally influences Y doesn't imply that changing D will result in a change in Y in all situations. (Smoldering cigarette stubs cause forest fires, but only during droughts. By the same token, droughts cause forest fires, but these need a spark to get started.)

1.3 Visualising causality: directed acyclic graphs (DAGS)

1.3.1 Why?

Research would be pretty easy if you could safely conclude that a causal link existed between two variables any time you observed an association between them. Fortunately for teachers of methodology courses who'd be on the dole otherwise, this isn't the case. But simply parrotting back *Correlation is not causation* isn't too helpful. To help us figure out how associations between two variables can arise in the absence of a causal link between them, we turn to **directed acylic graphs** (DAGS).

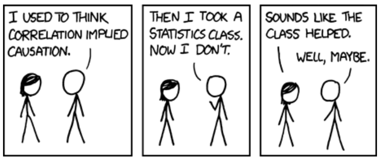

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Figure 1.1: Source: https://xkcd.com/552.

Graphs are mathematical objects in which nodes can be connected by edges. In directed graphs, these edges point from one node to another, i.e., they're arrows. If, in a directed graph, it is

[2] You'll also often see the words 'association' and 'correlation' used interchangeably. I prefer to use 'association' as the hypernym and reserve 'correlation' for a specific type of association. See Chapter 8.

impossible to start from some node and end up in the same node by following edges, then that graph is also acyclic. The examples below will make this a lot clearer. Graphs are studied in mathematics and computer science for sundry purposes; here, we will use DAGs as a tool for visually representing the causal links between the factors at play in a study.

As we'll see below, when DAGs are used for the purpose of representing causal links, they are subject to a number of rules that may appear cumbersome at first. However, when DAGs are properly specified, they allow researchers to figure out which factors they *should* control for, which factors they *can but needn't* control for and which factors they *must not* control for. Moreover, DAGs are useful for learning how associations in empirical data can occur both in the presence and in the absence of causal links between the variables of interest.

1.3.2 Some examples

Before laying down the rules for drawing DAGs, let's look at a couple of possible DAGs for the Andalusian CLIL study.

Figure 1.2 is the simplest of DAGs. It represents the assumption that there is a direct causal influence from the **treatment** variable (CLIL) on the **outcome** variable. These variables are represented by nodes. There exists a directed edge (i.e., an arrow) between them that shows the assumed direction of the causal link.

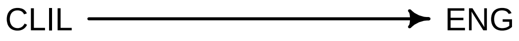


Figure 1.2: DAG representing a causal influence of CLIL on English proficiency (ENG).

The pupils' English proficiency won't be affected by their taking CLIL classes or not *alone* but by a host of other unobserved factors as well. In Figure 1.3, the unobserved factors are conveniently bundled and represented as 'U'. The U is circled to make it clear that these factors were not observed or measured. While this convention isn't universal, it's useful and we'll adopt it here.

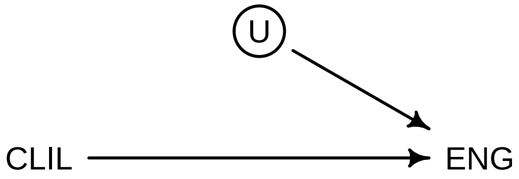


Figure 1.3: DAG representing a causal influence of CLIL and of unobserved factors on English proficiency.

Important: If we don't draw an arrow between U and CLIL, this means that we assume that there is *no* direct causal relationship between these two factors. But presumably, some unobserved factors will also account for why some pupils are enrolled in CLIL classes and others aren't; see Figure 1.4. As we'll discuss later, these unobserved factors, some of which may affect both the 'treatment' (CLIL) and the 'outcome' (English proficiency), **confound** the causal link of interest.

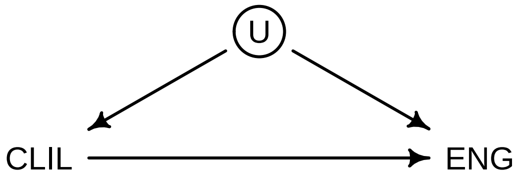


Figure 1.4: Unobserved factors as confounders (1).

Figure 1.5 also features the unobserved factors as possible confounders, but this time there is no assumed causal link between CLIL and ENG.

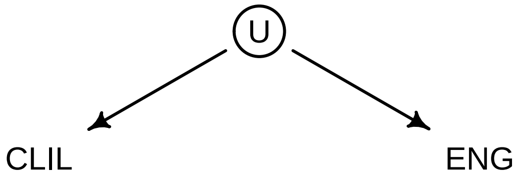


Figure 1.5: Unobserved factors as confounders (2).

1.3.3 *Rules for drawing* DAGS

1. The direction of the arrows shows the direction of the assumed causality (hence *directed*).
2. Bidirectional arrows are forbidden, i.e., no A ↔ B.
3. You're not allowed to draw graphs where you can end up at the same place where you started by just following arrows (hence *acyclic*). For instance, you're not allowed to draw a DAG like Figure 1.6.
   Mutual influencing factors can be represented in a DAG, however, but you need to break down the temporal structure. Figure 1.7 shows how you can break down the temporal structure implicit in Figure 1.6 to produce a legal DAG.

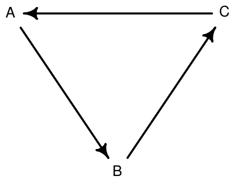


Figure 1.6: An illegal DAG: A, B and C are allowed to influence themselves.

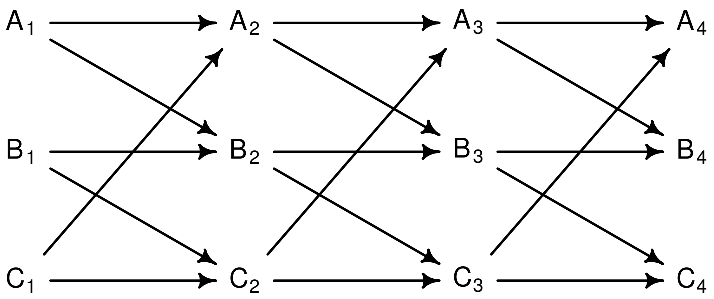


Figure 1.7: Reciprocal influences can be represented legally in a DAG if you break down the temporal structure. A_1, A_2, A_3 and A_4 represent the same variable measured at four points in time. The value of this variable at a given point in time is determined in part by its value at the previous point in time (e.g., A_3 is influenced by A_2) as well as by the value of another variable at the previous point in time (e.g., C_2 influences A_3).

4. Unobserved factors can, and often should, be drawn. By convention, we draw a circle around them to make it clear that they are not directly observed.

5. "DAGs insistently redirect the analyst's attention to justifying what arrows do not exist. Present arrows represent the analyst's ignorance. Missing arrows, by contrast, represent definitive claims of knowledge." (Elwert, 2013, p. 248)

6. A factor that isn't of interest and that only affects one factor already in the DAG and/or is affected by only one factor already in the DAG doesn't have to be drawn for you to be able to derive correct conclusions from the DAG. For instance, the U in Figure 1.3 doesn't have to be drawn (it only affects one factor that was already in the DAG). However, the U in Figure 1.4 *does* have to be drawn since it affects *two* factors that were already in the DAG. That said, it can be difficult to decide if a variable should be included in a DAG or not, and we shouldn't let perfect be the enemy of good.

Exercise 1.5. Draw a DAG that represents the belief that Danes' attitudes towards Swedish and their understanding of Swedish causally affect each other (i.e., the more the like it, the better they understand it, which leads to their liking it even better). ♢

Exercise 1.6. Draw a DAG that represents the belief that Danes who like Swedish seek out more contact with Swedish (e.g., by watching Swedish television), which leads to their understanding it better, which in turn leads to their seeking out even more contact with Swedish, and so on. ♢

1.3.4 Chains, forks and inverted forks

A DAG that is drawn by following the rules specified above is always built up out of at most three types of building blocks: chains, forks, and inverted forks.

Chains A chain is a sequence of causal links. In Figure 1.8, A → B → C → D forms a causal chain. Note that causality doesn't flow 'upstream' against the direction of the arrows, so there is no causal chain from D back to A.

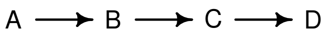


Figure 1.8: A chain.

Chains may transmit genuine causal influences, that is, altering the values of (say) A may bring about a change in some values in (say) D. In other words, A may causally affect D, albeit indirectly through B and C. Since the causality is directional, altering the values of D won't bring about any changes in the values of A, B or C.

Moreover, **chains may induce associations between the variables involved**. Based on the DAG in Figure 1.8, we wouldn't be surprised to find some association between the values of A, B, C and D. The DAG doesn't tell us what this association will look like, but we'll encounter some common forms of association in the weeks to come.

Note that it is possible that changes in A are not reflected in changes in D downstream, for instance because the effect that A has on B is quite small and only large changes in B affect C. This is why I wrote that changes in A *may* (rather than *will*) bring about changes in D.

If, for whatever reason, you want to prevent a chain from transmitting associations between two variables, the paths between these variables have to be **blocked** somewhere. This is achieved by **controlling** for one (or several) of the variables along the path. We'll discuss how you can control for a variable in more detail in the weeks to come, but a conceptually easy (if often practically arduous) way is to ensure that only people, words, etc. with the same value on that variable are included in the study. For instance, if for some reason you need to control for eye colour, you could include only green-eyed people in your study.

Forks When a single factor causally affects two or more other factors, a fork is formed; see Figure 1.9. In this example, A causally influences both B and C.

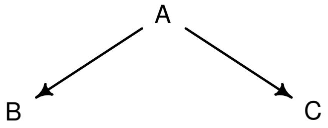


Figure 1.9: A fork.

Forks themselves don't transmit causal influences between the prongs, that is, altering the values of B won't change the values of C and vice versa: Causality doesn't travel upstream. If you want to represent a causal link between B and C, you have to add it to the DAG.

Importantly, **forks may induce associations between the factors at the prongs**: Based on the DAG in Figure 1.9, we wouldn't be surprised to find some association between the values of B and C. This is not because of a causal link between them but because A influences both of them. A is also referred to as a **confounding variable** or **confounder**.

To better appreciate the fact that causal forks can give rise to associations between the variables at the prongs, consider the fictitious example in Table 1.1. Here, A causally influences both B and C, and both B and C are additionally influenced by separate factors (U_B and U_C). The causal factors A, U_B and U_C can each take on two values (0, 1), and the outcomes of B and C are determined by simple equations. We assume that the probability of observing the values in a particular row is the same for all rows, i.e., that it is 1/8.

Table 1.1: Illustration of how a causal fork can give rise to associations between the variables at the prongs.

A	U_B	U_C	$B := A + U_B$	$C := A + U_C$
0	0	0	0	0
0	0	1	0	1
0	1	0	1	0
0	1	1	1	1
1	0	0	1	1
1	0	1	1	2
1	1	0	2	1
1	1	1	2	2

Taking a closer look at this table, we see that B and C are associated: The overall probability that B is at least equal to 1 is

$6/8 = 75\%$. But if you already know that an observation's value for C is 2, then you can be absolutely confident that its B value is at least 1. By the same token, if you know that its C value is 0, you'd be less confident about this guess:[3]

- $\mathbb{P}(B \geq 1 \mid C = 0) = 1/2$.
- $\mathbb{P}(B \geq 1 \mid C = 1) = 3/4$.
- $\mathbb{P}(B \geq 1 \mid C = 2) = 2/2$.

If you want to prevent a fork from transmitting an association between the variables at the prongs, you can control for the confounder or otherwise block the path on which the confounder lies. To appreciate this fact, again consider Table 1.1. We've already established that $\mathbb{P}(B \geq 1 \mid C = 0) \neq \mathbb{P}(B \geq 1 \mid C = 1)$. But once we 'control for' A by fixing it at a specific value (e.g., $A = 0$), we find that the probability of observing $B \geq 1$ doesn't depend on C any more.

- $\mathbb{P}(B \geq 1 \mid C = 0, A = 0) = 1/2$.
- $\mathbb{P}(B \geq 1 \mid C = 1, A = 0) = 1/2$.

Similarly, we could fix A at 1 and vary C and observe the same phenomenon:[4]

- $\mathbb{P}(B \geq 2 \mid C = 1, A = 1) = 1/2$.
- $\mathbb{P}(B \geq 2 \mid C = 2, A = 1) = 1/2$.

A silly example may be helpful to internalise the concept: While I don't have the numbers handy, I'm confident that there is some positive association between the number of drownings in the Aare and the daily revenue of Bernese ice-cream vendors. Why?

Inverted forks Figure 1.10 shows an inverted fork where two variables both influence a third one. The 'handle' of an inverted fork is called a **collider** since the two causal arrows clash into each other in A.

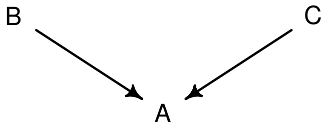


Figure 1.10: An inverted fork.

Inverted forks don't transmit causal influences between the variables at the prongs, that is, there is no causal link between B and C (causality doesn't travel upstream). The intriguing thing about inverted forks is this, though: When the collider (i.e., A) is *not* controlled for, the variables at the prongs remain unassociated. However, **controlling for the collider may induce an association between the variables at the prongs even in the absence of a causal link between them**. Controlling for a **descendant** of a collider may likewise induce an association between the variables at the prongs.[5]

The effects of controlling for a collider are not intuitive, so let's consider an example.

[3] $\mathbb{P}(B \geq 1 \mid C = 0)$ reads as 'the probability that B will be at least 1 when C equals 0.'

[4] We can't fix A at 1 and evaluate this probability at $C = 0$ for the simple reason that there's no row in the table with $A = 1$ and $C = 0$.

[5] A factor's descendants are the factors that are causally affected by it. A factor's **ancestors** are the factors that it is affected by.

University teachers can testify that there is some negative association between their students' intelligence and their diligence. This doesn't mean that the most intelligent students are *all* lazy and none of the most diligent students are particularly clever—just that there is some tendency for the most intelligent students to be less hard-working than the less clever ones. There is a simple and plausible causal explanation for this association: The most intelligent students quickly figure out that they don't need to work as hard in order to obtain their degree, so they shift down a gear.

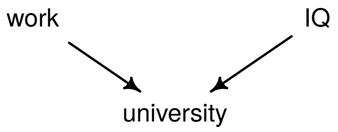


Figure 1.11: To the extent that diligence (work) and intelligence (IQ) both determine if someone gets into university, some association between these two factors will be found if we only look at university students.

But there is an equally plausible if less simple explanation: by only looking at university students, we've controlled for a collider without realising it; see Figure 1.11. Even if diligence and intelligence are completely unassociated in the human population, they are bound to be associated if we only look at university students. Figure 1.12 illustrates why: If we consider the population as a whole, it's possible that there is no (or hardly any) association between diligence and intelligence (left panel): If we know a person's degree of diligence, we can't make a more educated guess as to their intelligence than if we don't. But if we only consider university students (filled circles in the right panel), we're bound to find a negative association between diligence and intelligence: If we know that a university student is pretty lazy, we also know that they need to be pretty intelligent—otherwise they couldn't have made it into university.

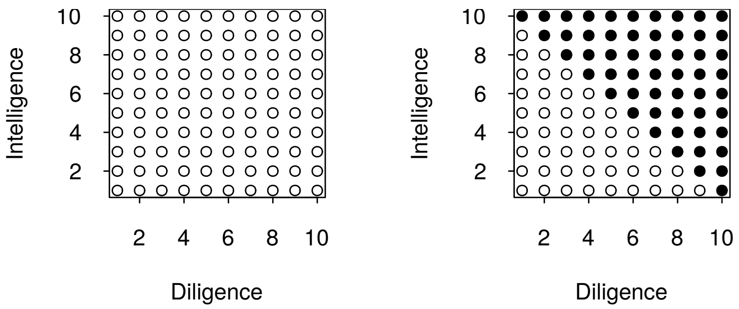


Figure 1.12: Collider bias in action: If you only look at the filled (or at the unfilled) circles, you'll discover an association between diligence and intelligence, even though there is no causal link between them.

By only looking at university students, we've unwittingly controlled for a collider, which by itself can explain the negative association between diligence and intelligence observed among university students. This doesn't mean that our first causal explanation is necessarily wrong, but it does illustrate that there is a non-obvious but plausible additional explanation that we need to reckon with. Note also that both explanations can simultaneously be correct: There may be some (negative) causal influence of intelligence on diligence, but by only looking at university students, we would then end up overstating the strength of this causal effect.

In Figure 1.12, we've assumed—for ease of exposition—that there is a perfect deterministic relationship between diligence and intelligence on the one hand and university enrolment on the other hand (viz., if the sum of both scores is above 10, enrolment is granted).

In reality, this relationship won't be perfect (some highly intelligent and highly diligent people don't go to university), but even so, controlling for (or 'conditioning on') a collider can produce associations between two factors in the absence of a causal link between them.

Other fairly common examples of this collider bias are only superficially different:

- There is a negative association between how easily accessible a restaurant is from a tourist resort and how good the food is. Come up with an explanation that does not assume any direct or indirect causal influence of food quality on location or vice versa.

- People with a highly active dating life sometimes complain that their hottest dates tend to be comparatively boring. Come up with an explanation that does not assume any direct or indirect causal influence of attractiveness on interestingness.

> In sum, unbroken chains both transmit causality and induce associations; forks induce associations without causality unless measures are taken (e.g., controlling for the confounder); and inverted forks induce associations without causality if the collider (or one of its descendants) is controlled for.

Exercise 1.7. Consider the DAG in Figure 1.13.

(a) Can A causally affect F?

(b) Can C causally affect D?

(c) Can there be an association between C and D if no factors are controlled for? Why (not)?

(d) Can there be an association between C and D if E is controlled for? Why (not)?

(e) Can there be an association between C and D if F is controlled for? Why (not)?

(f) Can there be an association between C and D if A is controlled for? Why (not)?

(g) Can there be an association between C and D if B is controlled for? Why (not)?

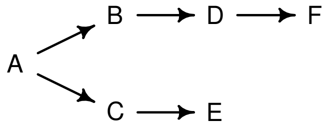


Figure 1.13: DAG for Exercise 1.7.

◇

Exercise 1.8. Consider the DAG in Figure 1.14.

(a) Can A causally affect F?

(b) Can A causally affect E?

(c) Can there be an association between A and E? Why (not)?

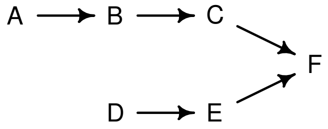


Figure 1.14: DAG for Exercise 1.8.

(d) Can there be an association between A and E if F is controlled for? Why (not)?

(e) Can there be an association between B and D if no factors are controlled for? Why (not)?

(f) Can there be an association between B and D if A is controlled for? Why (not)?

(g) Can there be an association between B and D if F is controlled for? Why (not)?

(h) Can there be an association between B and D if C and F are controlled for? Why (not)?

(i) Can there be an association between B and D if E and F are controlled for? Why (not)?

◊

Exercise 1.9. Consider the DAG in Figure 1.15.

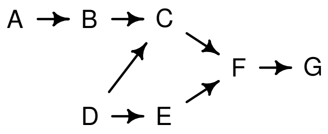


Figure 1.15: DAG for Exercise 1.9.

(a) Can A causally affect D?

(b) Can there be an association between A and D if no factor is controlled for? If so, via which path?

(c) Can there be an association between A and D if C is controlled for? If so, via which path?

(d) Can there be an association between A and D if F is controlled for? If so, via which paths (plural!)?

(e) Can there be an association between A and D if G is controlled for? If so, via which paths (plural!)?

(f) Can there be an association between A and D if B and C are controlled for? If so, via which path?

(g) Can C causally affect E?

(h) Can there be an association between C and E if no factor is controlled for? If so, via which path?

(i) Can there be an association between B and E if no factor is controlled for? If so, via which path?

(j) Can there be an association between B and E if C is controlled for? If so, via which path?

(k) Can there be an association between B and E if D is controlled for? If so, via which path?

(l) Can there be an association between B and E if D and F are controlled for? If so, via which path?

(m) Can there be an association between B and E if D and G are controlled for? If so, via which path?

◊

1.4 *Optional: Further reading*

Rohrer (2018) is an accessible introduction to DAGS. I don't recommended you read it right away, though, but save it in case you need a refresher from a different source in a couple of months or years.

Exercise 1.10 (Reading assignment)**.** Read Johnson (2013) with an eye towards explaining the following terms and concepts in a manner that you find intelligible by providing your own definition, clarifying example or illustration.

(a) Continuous vs. categorical variables (pp. 289–290 in Johnson, 2013)

(b) Histogram (pp. 292–293)

(c) Bimodal distribution (p. 292)

(d) Outliers (p. 292)

(e) Normal distribution (pp. 293–294)

(f) Arithmetic mean vs. median vs. mode (pp. 295–296)

(g) The effect of outliers on the mean and median (p. 297)

(h) Quantile, percentile, and quartile (p. 298)

(i) Standard deviation and variance (p. 299)

(j) Left- and right-skewed distributions (p. 301)

(k) Ordinal vs. nominal variables (p. 307)

(l) Bar chart (pp. 307–308)

(m) Contingency table (p. 311) ♢

2
Constructing a control group

2.1 A made-up example

Imagine that a new self-learning method for fostering Danish reading skills in speakers of German has been developed. You're tasked with finding out if this new method works better than the old one.

First attempt You find four students of German philology who want to learn Danish. You ask them to work autonomously with the new learning method half an hour a day for three weeks. After three weeks, you give them an article from a Danish newspaper, which they are to summarise orally in German. Two raters judge these summaries at their own discretion (20-point scale); the mean of the two ratings per learner counts as their reading comprehension score. The average group score is 11/20.
What can you conclude from this study?

One of several problems with this study is that there is no baseline against which to compare the participants' average result: We don't know whether 11/20 indicates that the new learning method works better or worse than the old one, or whether the old and new learning method are roughly equally effective. So we need a comparison or **control group** of people that didn't take part in the intervention.

Second attempt You convince four law students to also take part in the study. They're asked to work with the old learning method half an hour a day for three weeks. Then they take the same test as the German philology students. Their group mean is 8/20.

2.2 *Critical questions*

The second attempt outlined above also falls short on a number of criteria. There are a couple of critical questions we can ask, and slightly modified versions of these questions can be asked about studies in general.

Internal validity Can the difference in test scores between the two groups be ascribed to the difference in learning methods, or do alternative explanations for it present themselves?

External validity Does the finding apply only to the present **sample** or also to a larger **population**? To what population, exactly?

Ecological validity (Especially for applied research.) To what extent are the findings applicable to the world outside of the lab (e.g., teaching, policy)?

Internal reliability (a) Confronted with the same data, would other researchers draw similar conclusions? (The overall results may leave room for interpretation.) (b) Are the measurements consistent? For instance, would different observers agree on the measurements? (The raw data may leave room for interpretation.)

External reliability Can the results of this study be confirmed in an independent **replication**, that is, in a new but otherwise similar study?

The definitions of different types of validity and reliability vary from source to source. The labels aren't too important; the questions behind them are.

Questions concerning validity and reliability can rarely be answered with a clear 'present' or 'absent'. But our second attempt outlined above is deficient in both respects. Discuss a few problems.

Some relevant terminology:

Confounding variable See Chapter 1.

Inter-rater reliability The extent to which different raters would score the observations similarly.

Intra-rater reliability The extent to which the *same* raters would score the observations similarly on a different occasion.

Anticipate and resolve problems related to lacking validity and reliability *before* collecting the data. This often involves making compromises or coming to the realisation that you can't satisfactorily answer all your questions in a single study. Do *not* assume that some statistical method will solve your problems.

Depending on your goals, some types of validity or reliability may not be as important as others. For instance, for most studies in psycholinguistics, university students are recruited as participants, and their results don't necessarily generalise to the population at large. But the purpose of these studies is often to demonstrate that some experimental manipulation *can* affect language use and processing, not that it will yield the same exact effect for everyone. From this perspective, these studies' lack of external validity isn't too damning (Mook, 1983).

2.3 *Increasing internal validity through randomisation*

Our first priority is to maximise the study's internal validity, that is, we want to maximise the chances that any association we find the data is due to the factor of interest. Confounding in particular represents a substantial threat to internal validity: As we've seen in Chapter 1, confounding variables induce associations between the variables of interest even in the absence of a causal link between them. Moreover, even if a causal link does exist between the variables of interest, confounding variables can **bias**[1] the association between them: The association may systematically under- or overestimate the strength of the causal link. Keeping confounding in check is therefore key.

Your first inclination may be to try to ensure that the intervention and control groups are identical in all respects save for the treatment itself. That way, any differences in the outcome variable can't be explained by confounding due to pre-existing differences between the groups. However, it is usually impossible to assign a fixed number of participants to two groups in such a way that these groups are identical in all respects even in the utterly unrealistic case where all the relevant information is available beforehand. Clearly, it's entirely impossible to do so when not all of the relevant information is available beforehand.

The solution is to assign the participants (or whatever your units of observation are) to the study's conditions **at random**, i.e., to deliberately leave the allocation up to chance and chance alone. The DAGs in Figure 2.1 on the next page show what such randomisation achieves. When the participants themselves (or their parents, or their circumstances, etc.) determine which condition they end up in (X), confounding is a genuine concern (left). However, when we assign the participants to the conditions at random, we *know* that there is no systematic link between pre-existing characteristics (U) and X, let alone a causal one. That is, randomisation prevents any causal arrows from entering X (right)! The result of this is that the non-causal path between X and Y (via U) is broken and that the X-Y relationship is no longer confounded by U.

Studies in which the participants (or whatever the units of observation are) are randomly assigned are called **true experiments**.Random allocation by itself doesn't guarantee that the results

[1] The term *bias* refers to a systematic distortion of the results, e.g., due to confounding variables. A single unbiased study isn't guaranteed to estimate the size of the causal effect correctly, but roughy speaking, if we were to run the same study lots of times, the under- and overestimates would cancel each other out. If the under- and overestimates don't cancel each other out, then the study is biased.

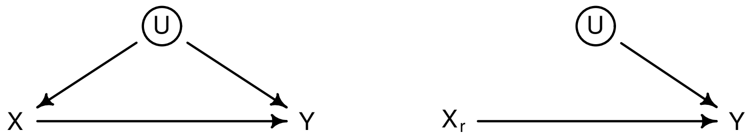


Figure 2.1: *Left:* The X-Y relationship is confounded by U: there are two paths from X to Y, but only one causal one. *Right:* Randomising the values of X prevents arrows from U entering X, which effectively closes the non-causal path via the confounder.

of the experiment can be trusted or interpreted at face value, but it does eliminate one common threat to the study's internal validity: confounding.

Randomise wherever possible – unless you have a very good reason not to (see weeks to come)!

2.3.1 *Why experiments?*

1. "Experiments allow us to set up a **direct comparison** between the treatments of interest.

2. "We can design experiments to **minimize any bias** in the comparison. [especially randomisation]

3. "We can design experiments so that the **error** in the comparison is **small**. [see weeks to come]

4. "Most important, we are **in control** of experiments, and having that control allows us to make stronger inferences about the nature of differences that we see in the experiment. Specifically, we may make **inferences about causation**." (Oehlert, 2010, p. 2, my emphasis)

2.3.2 *What does randomisation do?*

1. "Randomization balances the population on average."

2. "The beauty of randomization is that it helps prevent confounding, *even for factors that we do not know are important*." (Oehlert, 2010, p. 15, my emphasis)

We've already discussed the second point, but the first point warrants some explanation. Let's say that you have ten participants and you know both their sex and their IQ (Figure 2.2 on the facing page). If you randomly assign these participants to two conditions with five participants each, you may end up with one of the six allocations shown in Figure 2.3—or any of the 246 others.[2] Note that in none of them, the intervention and control groups are perfectly balanced with respect to both IQ and sex. So randomisation clearly does not generate balanced groups in any particular study. However, each participant is as likely to end up in the intervention group as they are to end up in the control group, so *on average*—across all 252 possible random allocations—sex, IQ, as well as all unmeasured variables, are balanced between the two groups. For our present purposes, this means that randomisation is

[2] There are $\binom{10}{5} = \frac{10!}{5! \cdot (10-5)!} = 252$ different ways to split up ten people into two groups of five. '5!' (read: 'five factorial') means $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$, and '$\binom{10}{5}$' is read as '10 choose 5'.

an equaliser: the result may not be two perfectly equal groups, but at least one group isn't systematically given an advantage relative to the other. As we'll see in Chapter 4, randomisation also justifies the use of some common statistical procedures.

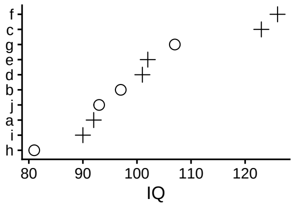


Figure 2.2: Ten participants sign up for a study. You measure their IQ and you also know their sex (represented here using circles and crosses).

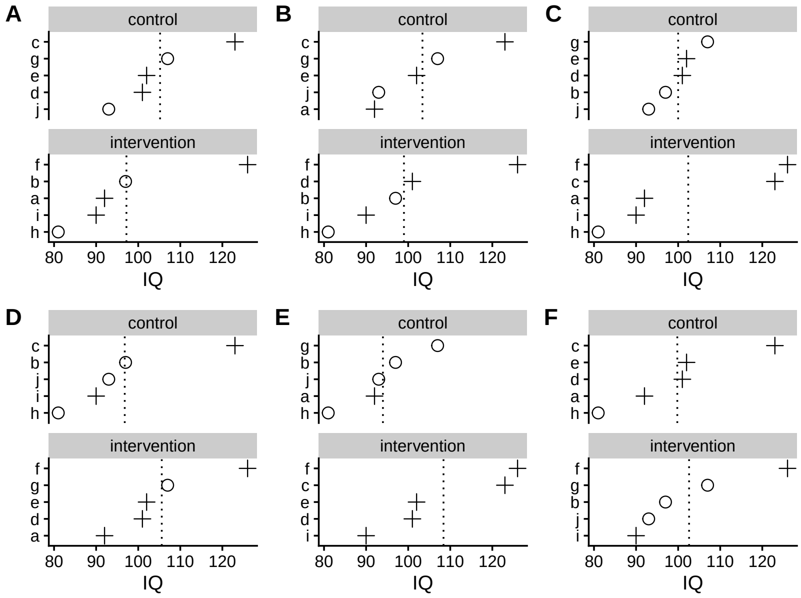


Figure 2.3: Six possible random assignments (out of 252) of the ten participants from Figure 2.2. The dotted vertical lines show the mean IQ in each group.

Exercise 2.1. For each description, decide whether the participants were randomly assigned to the experiment's conditions and, if not, explain how the lack of randomisation could result in confounding.

(a) 60 participants trickle into the lab. The first 30 are assigned to the experimental condition, the final 30 are assigned to the control group.

(b) Experiment with a school class: Pupils whose last name starts with a letter between A and K are assigned to the control group, the others to the experimental group.

(c) Participants come to the lab one by one. For each participant, the researcher throws a dice. If the dice comes up 1, 2, or 3, the participant is assigned to the experimental condition; if it comes up 4, 5, or 6, the participant is assigned to the control condition. After four weeks, no more participants sign up. The control group consists of 17 participants; the experimental group of 12.

(d) To investigate the effects of bilingualism on children's cognitive development, 20 bilingual 4-year-olds (10 girls, 10 boys) are recruited. 20 monolingual 4-year-olds (10 girls, 10 boys) serve as the control group.

(e) 32 participants sign up for an experiment. The researcher enters their names into `http://www.random.org/lists/`, clicks `Randomize` and assigns the first 16 to the control group and the others to the experimental group. ♢

> 'Random' does not mean 'haphazard', 'arbitrary' or 'at the researcher's whim'.

2.3.3 *How to randomise?*

When collecting data using computers Have the computer randomly assign the participants to the conditions without your involvement. Programmes for running experiments such as OpenSesame (`https://osdoc.cogsci.nl/`), PsychoPy (`https://www.psychopy.org/`) or jsPsych (`https://www.jspsych.org/`) all contain functions for allocating participants randomly.

When the data collection does not take place at the computer and you know who'll be participanting beforehand Randomise the list of participants using `https://www.random.org/`. Assign the first half of the list to the experimental condition and the second half to the control condition.

This procedure is known as **complete randomisation**. It guarantees that the number of experimental units is the same in each condition (or at most one off if the number of units isn't divisible by the number of conditions).

When the data collection does not take place at the computer and you don't know who'll be participating beforehand Randomly assign each participant individually and with the same probability to a condition as they sign up. This procedure is known as **simple randomisation**. In contrast to complete randomisation, you're not guaranteed to end up with an equal number of units in each condition. This is usually of little concern, and in fact, simple randomisation arguably reduces the potential for the researchers' biases to affect the study's results (Kahan et al., 2015). Importantly, **there is nothing wrong with having unequal sample sizes.**[3]

> Humans make for poor randomisation devices. Always randomise mechanically (preferably with a computer).

Exercise 2.2. For each description, decide if the study is a true experiment.

[3] See blog entry *Causes and consequences of unequal sample sizes*.

(a) Eight Swiss speakers of German indicate how beautiful they find the French language on a 7-point scale. Additionally, they all record a text in French. In a ‘perception experiment’, 20 native speakers rate all recordings on a 5-point scale from ‘very strong foreign accent’ till ‘no foreign accent whatsoever’. The question is whether the speakers’ attitudes are related to the strength of their accent in French (Kolly, 2011).

(b) ”This study presents the first experimental evidence that singing can facilitate short-term paired-associate phrase learning in an unfamiliar language (Hungarian). Sixty adult participants were randomly assigned to one of three ”listen-and-repeat” learning conditions: speaking, rhythmic speaking, or singing.” After 15 minutes of learning, the learners’ Hungarian skills are tested and compared between the three conditions (Ludke et al., 2014).

(c) ”The possible advantage of bilingual children over monolinguals in analyzing word meaning from verbal context was examined. The subjects were 40 third-grade children (20 bilingual and 20 monolingual) ... The two groups of participants were compared on their performance on a standardized test of receptive vocabulary and an experimental measure of word meanings, the Word–Context Test.” (Marinova-Todd, 2011)

(d) ”The present paper considers the perceived emotional weight of the phrase *I love you* in multilinguals’ different languages. The sample consists of 1459 adult multilinguals speaking a total of 77 different first languages. They filled out an on-line questionnaire with open and closed questions linked to language behavior and emotions. Feedback on the open question related to perceived emotional weight of the phrase *I love you* in the multilinguals’ different languages was recoded in three categories: it being strongest in (1) the first language (L1), (2) the first language and a foreign language, and (3) a foreign language (LX) ... Statistical analyses revealed that the perception of weight of the phrase *I love you* was associated with self-perceived language dominance, context of acquisition of the L2, age of onset of learning the L2, degree of socialization in the L2, nature of the network of interlocutors in the L2, and self-perceived oral proficiency in the L2.” (Dewaele, 2008) ♢

The word ’experiment’ can be used in a stricter or in a looser sense. The mere fact that a study is referred to as an ’experiment’ does *not* mean that it’s a *true experiment* (control group + randomisation): the use of the label doesn’t automatically imply that confounding has been taken care of.

Most quantitative studies in our research area aren’t experiments in the strict sense.

Exercise 2.3 (Reading assignment)**.** Read Ludke et al. (2014) in light of the questions below and briefly answer them. As you're reading the results section, focus on the descriptive statistics and the graphs; you don't have to bother with the statistical tests for now.

(a) What is this study's most important research question or its most important aim?

(b) Briefly describe this study's design. Is this study a 'true experiment'?

(c) How did the researchers try to ensure that any differences between the conditions could be attributed to differences between speaking, rhythmic speaking and singing rather than to other factors?

(d) "Digital audio recordings were made during each experimental session." (p. 46, 2nd column, last paragraph) Why?

(e) What does this mean: "Measures of participants' mood, background experience, and abilities in music and language were also administered in order to check that the randomly assigned groups were matched for these factors." (p. 43, 2nd column, last paragraph) ♢

3
Alternative explanations

3.1 The roles of variables in research

Some common terminology:

Dependent variable or *outcome variable*.

Independent variable or *predictor variable*. In experiments, such variables are 'manipulated' by the researchers.

Control variable. Additional variable that was collected as it may be related to the *outcome*. We'll discuss the usefulness of control variables later.

3.2 Alternative explanations for results

In Chapter 2, we focused on the threat that confounding poses to a study's internal validity and how this threat can be neutralised using randomisation. We saw that randomised ('true') experiments (probabilistically) negate the influence of confounding variables on the results: one group isn't systematically given an advantage compared to the other (e.g., higher motivation, greater affinity with a topic etc.). This increases the study's internal validy, but:

> Even if confounding variables are taken into account, other systematic factors may give rise to a spurious difference between the experiment's conditions or may mask an existing effect of the conditions.

3.3 Explanation 1: Expectancy effects

Perhaps the researchers or their assistants (subconsciously) nudged the data in the hypothesised direction. This can happen even when the measurements seem perfectly objective. For instance, when you're counting the number of syllables in a snippet of speech, there are bound to be a number of close decisions (Does German *haben* [ha(b)m] have one or two syllables?). This isn't too big a problem in itself, but it does become a cause for concern if you tend to reach different decisions depending on which condition the participant was assigned to.

Relatedly, it's possible that the participants want to help (or thwart) the researchers achieve what they think are the researchers' goals. In this case, differences in the outcome variable between the conditions may arise not because of the intervention itself but because of unwanted changes in the participants' behaviour. Such changes in behaviour needn't come about consciously.

Expectancy effects Both on the part of the participants (e.g., *placebo* effect) or on the part of the researchers.

Single-blind experiment Typically used to describe that the participants don't know which condition they're assigned to.

Double-blind experiment If neither the participants nor the researchers themselves (at the time of collecting and preparing the data) know which condition the participants were assigned to.

Blinding isn't always possible, and it may be immediately obvious to the participants what the intervention entails. But in studies with raters, it's usually easy to prevent them from knowing which condition the participants were assigned to.

3.4 *Explanation 2: Failed manipulation*

A second class of alternative explanations is that the experiment didn't run quite as the researchers expected it to. For instance, the participants may have misunderstood, or failed to act on, the instructions, or the script used to run the experiment could contain a crucial coding error.

Manipulation checks Example 1: Ludke et al. (2014) wanted to find out if foreign-language phrases are more easily learnt if the learners practice them while singing or speaking rhythmically. Their experimental manipulation involved asking their participants to practice Hungarian phrases while singing or speaking rhythmically. In order to verify whether they indeed did as they were asked, they recorded their participants.

Example 2: Lardiere (2006) had her participant judge L2 sentences for their grammaticality. To ensure that the participant rejected sentences for the (syntactic) reason intended by the researcher, she was also asked to correct any sentences she rejected. The researcher found out that the participant rejected a fair number of syntactically correct sentences, but that she did so for stylistic (rather than strictly syntactic) reasons. The researcher then (correctly) didn't draw the conclusion that the participant's syntactic knowledge was incomplete.

Satisficing Sometimes, participants don't really pay any attention to the stimuli or to the instructions. For instance, questionnaire respondents may answer in a specific pattern (e.g., ABCDEDCBA...) rather than give their mind to each question. Figure 3.1 provides another example of satisficing.

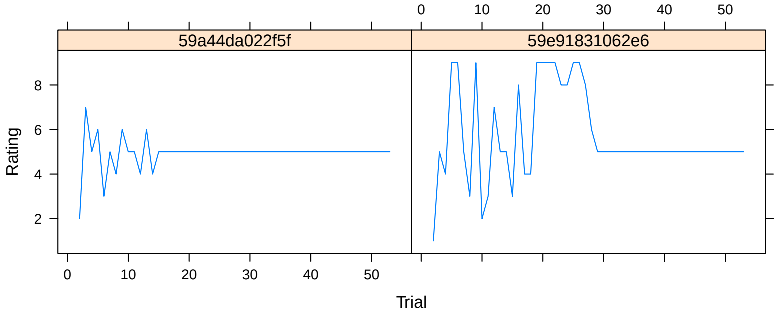


Figure 3.1: A large number of raters were asked to each rate about 50 short texts on a 9-point scale. These two clearly lost interest at some point (Vanhove, 2017).

If you want to run a study online or at the computer, check out Oppenheimer et al. (2009) for a neat and unintrusive way to find out if your participants read the instructions.

Positive control Does the intervention yield an effect in cases where it *should* (with near-certainty) yield an effect? If not, then the experiment may have been carried out suboptimally.

Example: In L2 research, the task given to the L2 speakers is sometimes also given to a group of L1 speakers to make sure that the latter can complete it.

The term *negative control* refers to traditional control groups (of which we know that they shouldn't show an effect of the intervention).

Pilot study See classes on questionnaires. Some goals of pilot studies are to make sure that the participants understand, and act on, the instructions, identify any remaining glitches in the experimental software and (if relevant) check if the responses obtained can be coded satisfactorily.

3.5 *Explanation 3: Chance*

A third important possible non-causal explanation for one's results is that they're due to chance. The entire next chapter is devoted to attempt to get a handle on this explanation.

4 Inferential statistics 101

This chapter explains the basic logic behind p-value-based inferential statistics. It does so by explicitly linking the computation of p-values to the random assignment of participants to conditions in experimental research. If you have ever taken an introductory statistics class, chances are p-values were explained to you in a different fashion, presumably by making assumptions about how the observations in the sample were sampled from a larger population and by making reference to the Central Limit Theorem. For the explanation in this chapter, however, we're going to take a different tack and we will ignore the sampling method and the larger population. Instead, we're going to leverage what we know about how the observations, once sampled, were *assigned* to the different conditions of an experiment. The advantages of this approach are that it connects the design of a study more explicitly to the analysis of its data and that it is less math-intensive while permitting one to illustrate several key concepts about inferential statistics.

The goal of this chapter is for you to *understand conceptually* what statistical tests attempt to achieve, not for you to be able to use them yourself. As a matter of personal opinion, statistical tests are overused (Vanhove, 2021). I think that, in your own research, your focus should be on describing your data (e.g., by means of appropriate graphs) rather than running umpteen significance tests. Analysing data and running statistical tests are not synonymous.

4.1 An example: Does alcohol intake affect fluency in a second language?

Research question Does moderate alcohol consumption affect verbal fluency in an L2?

Method Ten students (L1 German, L2 English)[1] are randomly assigned to either the control or the experimental condition (five each); they don't know which condition they're assigned to. Participants in the experimental condition drink one pint of ordinary beer; those in the control condition drink one pint of alcohol-free beer.

Afterwards, they watch a video clip and relate what happens in it in English. This description is taped, and two independent raters

[1] Ten participants is obviously a very low number of participants, but it keeps things more tractable here.

who don't know which condition the participants were assigned to count the number of syllables uttered by the participants during the first minute. The mean of these two counts serves as the verbal fluency/speech rate variable.

Results The measured speech rates are shown in Figure 4.1. On average (mean), the participants in the *with alcohol* condition uttered 4.3 syllables/second, compared to 3.7 syllables/second in the *without alcohol* condition.

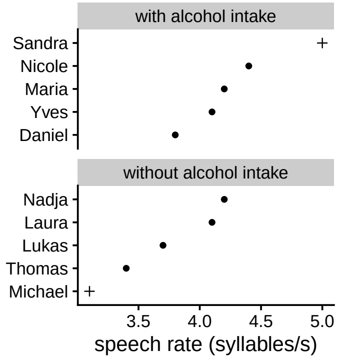


Figure 4.1: Individual results of a randomised experiment.

4.2 *The basic question in inferential statistics*

We have dealt with major threats to internal validity, viz., confounders (neutralised using randomisation) and expectancy effects (neutralised using double blinding). But there is another threat to internal validity that we need to keep in check: While we found a mean difference between the two conditions (4.3 vs. 3.7), this difference could have come about through *chance*. We are, then, faced with two types of accounts for this mean difference:

- The **null hypothesis** (or H_0): The difference between the means is due *only* to chance.
- The **alternative hypothesis** (or H_A): The difference between the means is due to chance *and* systematic factors.

Assuming the H_0 *is correct*, the participants' results aren't affected by the condition (alcohol vs. no alcohol) they were assigned to. For instance, Sandra was assigned to the *with alcohol* condition and her speech rate was measured to be 5.0. But had she been assigned to the *without alcohol* condition, her speech rate would also have been 5.0. Assuming the H_0 is correct, then, the difference in speech rate between the two conditions must be due solely to the random assignment of participants to conditions, due to which more fluent talkers ended up in the *with alcohol* condition. Another roll of the dice could have assigned Sandra to the control condition instead of Michael, and since under the H_0, the speech rate of neither is influenced by the condition, this would have produced a slower speech rate in the *with alcohol* condition than in the *without alcohol* one (3.9 vs. 4.1; see Figure 4.2).

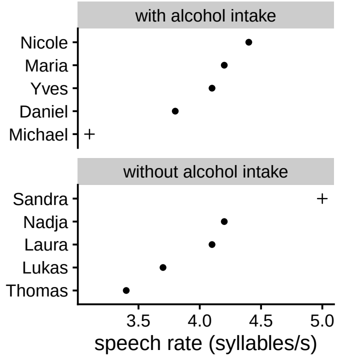


Figure 4.2: If only chance were at play, Michael's (3.1) and Sandra's (5.0) results would be unaffected by the experimental condition and the outcome might equally well have looked like this (swapping Michael and Sandra).

Frequentist inferential statistics seeks to quantify how surprising the results would be if we assume that only chance is at play. To do so, it attempts to answer the following key question: **How likely is it that a difference at least this large would've come about if chance alone were at play?**

If it's pretty unlikely that chance alone would give rise to at least the difference observed, then this can lead one to revisit the assumption that the results are due only to chance—perhaps some systematic factors are at play after all. By tradition, the threshold between 'pretty unlikely' and 'still too likely' is 5%, but there is nothing special about this number. If the result falls below this 5% threshold, the difference is said to be 'statistically significant'. This is just a phrase, however, and arguably a poorly chosen one: statistical 'significance' doesn't tell you anything about a result's practical or theoretical import.[2] Before discussing these and other misunderstandings about significance tests, let's see how you can compute how often you would observe a mean difference of at least $4.3 - 3.7 = 0.6$ if chance alone were at play.

4.3 *Testing the null hypothesis by exhausitive re-randomisation*

With 10 participants in two equal-sized groups, there were 252 possible assignments of participants to conditions, each of which was equally likely to occur. To see how easily a difference as least as large as the one observed (4.3 vs. 3.7) could occur due to random assignment alone, we can re-arrange the participants' speech rates into each of these 252 combinations and see for each combination what the difference between the *with* and *without alcohol* condition means is (Figure 4.3).

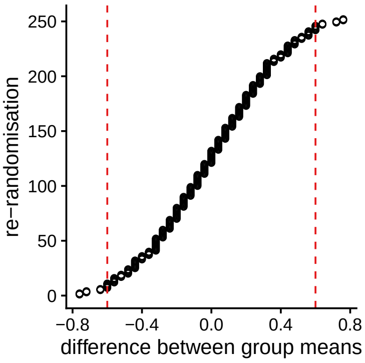


Figure 4.3: There are 252 different ways in which the 10 participants could have been split up into two groups. These are the differences between the means for all 252 possibilities.

In 22 out of 252 cases, the re-arrangement of participants to conditions produced a difference at least as large in magnitude (positive or negative) as the difference we actually observed. In other words, the probability with which we would observe a difference of at least 0.6 between the conditions if chance alone (random assignment) is at play is $\frac{22}{252} = 0.087$ (8.7%). This is the infamous **p-value**.

Since $p = 0.087 > 0.05$, one would typically conclude that a difference of 0.6 or more is still likely enough to occur under the null hypothesis of chance alone and that, hence, there is little need to revisit the assumption that the results may be due to chance alone. **Crucially, this doesn't mean that we have shown H_0 to be true.** It's just that H_0 would account reasonably well for these data.

4.4 *On 'rejecting' null hypotheses*

Researchers will often say that they 'reject' the null hypothesis in favour of the alternative hypothesis if $p \leq 0.05$. While this practice is subject to often heated debate (see McShane et al., 2019), it's important to realise that p can be ≤ 0.05 even if the null hypothesis is true,[3] and that $p > 0.05$ can occur even if the alternative hypothesis is true. Consequently, researchers who are in the business of 'rejecting' null hypotheses can make two types of errors, depending on whether H_0 or H_A is actually true.

	H_0 is actually correct	H_A is actually correct
$p > 0.05$	Fine—we didn't reject H_0	Wrong conclusion
$p \leq 0.05$	Wrong conclusion	Fine—we rejected H_0 in favour of H_A

Table 4.1: If you're in the business of rejecting null hypotheses, there are two types of errors you can make. Incorrectly rejecting the H_0 is commonly referred to as a **Type-I error**; incorrectly not rejecting the H_0 is referred to as a **Type-II error**.

Without additional information (e.g., in the form of converging

[2] From now on, avoid using the words 'significance' and 'significant' in their non-technical sense when writing about quantitative research.

[3] In theory, in fact, $p \leq \alpha$ in at most 100α% of the studies in which H_0 actually is true (for all $\alpha \in (0, 1)$). In practice, however, things aren't so simple. We'll return to this when discussing questionable research practices.

evidence from other studies or logical reasoning), we can't really know whether '$p \leq 0.05$' represents an error or a true finding. **(!)**

Note, furthermore, that the H_A stipulates that the results are due to a combination of chance and systematic factors. It doesn't stipulate *which* systematic factors, though. What we would like to conclude is that the systematic factor at play is our experimental manipulation, but expectancy effects, failed manipulations, confounding and collider bias are also systematic factors. What is more, the experimental manipulation may exert a systematic effect on the results, but for different reasons than we think they do.[4]

4.5 Analytical short-cuts

Exhausitive re-randomisation is cumbersome for larger samples[5] and more complex research designs; analytical short-cuts (e.g., the t-test, χ^2-test, ANOVA etc.) and their generalisations usually produce similar results and are used instead. In the context of experiments with random assignment, the p-values etc. that these procedures return have the essentially same interpretation and are subject to the same caveats as those above.

4.6 Statistical power

A study's statistical power is the probability with which its significance test will yield $p \leq 0.05$. In studies in which one group is compared to a different group, this probability depends on three factors (see Figure 4.4 on page 38):

1. The size of the difference in the outcome between the groups that the systematic factors cause. Even if they don't cause any difference, it is possible to obtain a statistically significant difference due to chance (see table above).

2. The number of observations.

3. The variability in the outcome variable within each group.

The precise numbers along the y-axis in Figure 4.4 aren't important; what's relevant is the direction and the shape of the curves.

Exercise 4.1. Take a look at Figure 4.4 and answer the following questions:

(a) How do the effect size, the number of observations and the within-group variability in the outcome affect the probability that a study will yield a statistically significant result?

(b) Other things equal, what yields a greater improvement in a study's power: 10 additional participants per group when each group already consists of 10 participants, or 20 additional participants per group, when each group already consists of 50 participants?

[4] For instance, a systematic difference between the *with* and *without alcohol* conditions needn't be due to alcohol intake per se but may be related to the taste of the beers in question instead. Or maybe alcohol increases speech rate—not because the participants become more fluent per se, but because they use simpler syntactic constructions that they can produce more quickly. In other studies, different theoretical explanations may account for any given finding—in addition to more mundane reasons such as confounding, expectancy effects and the like.

[5] How many ways are there to split up 40 participants into two equal-sized groups?

(c) How could researchers reduce the within-group variability in the outcome variable? ♢

Exercise 4.2. p-values are commonly misinterpreted. By way of preparation for the next exercise, answer the following questions.

(a) What, roughly, is the probability that, when you'll die, it'll be because a shark bit your head clean off?[6]

(b) What, roughly, is the probability that, when a shark bites your head clean off, you'll die?[7]

(c) Is 0.087 the probability that the null hypothesis in the alcohol example is correct? If not, then which probability exactly does this p-value of 0.087 refer to? ♢

Exercise 4.3. Consider the following vignette and some possible interpretations of the results reported in them. Decide for each interpretation if it follows logically from the vignette and the correct definition of the p-value. Explain your reasoning.

Vignette: In an experiment that was carried out and analysed rigorously, we find that the mean difference between the control and intervention groups amounts to 5 points on a 100-point scale. This difference is "statistically significant", with a p-value of 0.02.

(a) It's unlikely that we would have found a difference of 5 points or larger between both groups if the null hypothesis were indeed true. More precisely, this probability would have only been 2%.

(b) The null hypothesis is incorrect; the alternative hypothesis is correct.

(c) It's unlikely that the null hypothesis is indeed correct. More precisely, the probability that it is correct is only 2%.

(d) It's highly likely that the alternative hypothesis is correct. More precisely, the probability that it is correct is 98%.

(e) A new but similar study would likely yield a low p-value as well. More precisely, there is a 98% probability that such a study would yield a significant p-value (i.e., $p \leq 0.05$).

(f) If we concluded that the alternative hypothesis is correct, we would be wrong at most 2% of the time.

(g) If we concluded that the alternative hypothesis is correct, we would be wrong at most 5% of the time. ♢

Exercise 4.4. Consider the following vignette and some possible interpretations of the results reported in them. Decide for each interpretation if it follows logically from the vignette and the correct definition of the p-value. Explain your reasoning.

[6] In the notation of probability theory, you'd write this as $\mathbb{P}(\text{head bitten off by shark} \mid \text{dead})$.

[7] I.e., $\mathbb{P}(\text{dead} \mid \text{head bitten off by shark})$.

Vignette: In an experiment that was carried out and analysed rigorously, we find that the mean difference between the control and intervention groups amounts to 5 points on a 100-point scale. This difference is "not statistically significant", with a p-value of 0.64.

(a) It's pretty likely that we would have found a difference of 5 points or larger between both groups if the null hypothesis were indeed true. More precisely, this probability would have been 64%.

(b) The null hypothesis is correct; the alternative hypothesis is incorrect.

(c) It's pretty likely that the null hypothesis is indeed correct. More precisely, the probability that it is correct is 64%.

(d) It's fairly unlikely that the alternative hypothesis is correct. More precisely, the probability that it is correct is 36%.

(e) A new but similar study would likely yield a high p-value as well. More precisely, there is a 64% probability that such a study would yield a non-significant p-value (i.e., $p > 0.05$). ♢

Exercise 4.5. Assume that 10,000 experiments are carried out and analysed appropriately. Further assume that the null hypothesis is correct in all of these experiments. Which of the four histograms shown in Figure 4.5 on page 38 would the distribution of the 10,000 p-vaues resemble most closely? What if the alternative hypothesis were correct in all of the 10,000 experiments? ♢

Exercise 4.6. Imagine that in a given academic field, one thousand (1,000) experiments get conducted over the course of a year. Assume that in 80% of these experiments, the null hypothesis is correct, whereas in 20%, the alternative hypothesis is correct. Further assume that if the null hypothesis is correct, there is a 5% chance that a statistically significant difference will be observed between the experiment's conditions; if the alternative hypothesis is correct, there is a 60% chance of observing a statistically significant difference.

For the following questions, you may use the following probabilistic fact: If n attempts each succeed with probability p, then the expected number of successful attempts out of the n attempts is np.

(a) In how many of the 1,000 experiments can we expect that the null hypothesis will be incorrectly rejected?

(b) In how many of the 1,000 experiments can we expect that the null hypothesis will be correctly rejected?

(c) Now consider the total number of experiments in which we expect to reject the null hypothesis. In what percentage of these experiments will we have drawn the correct conclusion? ♢

Exercise 4.7. A research team is conducting an experiment in 30 school classes simultaneously. In each school class, the children are assigned to the control or experimental group using complete randomisation. That is, the researchers are in effect conducting 30 parallel and independent experiments.

Assume that in all of these classes, the null hypothesis is true. Further assume that, if the null hypothesis is true, there is exactly a 5% chance of observing a statistically significant difference between the control and experimental classes.

What is the probability of obtaining a statistically significant difference in at least one of the 30 experiments?

(Hint: What's the probability of observing 30 non-significant differences?) ◇

4.7 *Optional: Further reading*

The blog entries *Explaining key concepts using permutation tests* and *A purely graphical explanation of* p*-values* may be of some use. For an explanation in German, see Chapter 13 of my statistics booklet (available from `https://janhove.github.io`). Goodman (2008) discusses some common misinterpretations of p-values; his list is far from exhaustive.

Analysing quantitative data and running significance tests aren't synonymous; see my booklet as well as Winter (2019) for introductions to statistics for linguists that don't emphasise signficance testing. Nonetheless, quantitative research in the social sciences, including in applied linguistics, has developed something of a significance fetish, with authors inundating their readers with significance tests and p-values that they themselves don't seem to really understand while giving them little insight into what the data actually look like. For further lamentations and some suggestions, see Vanhove (2021).

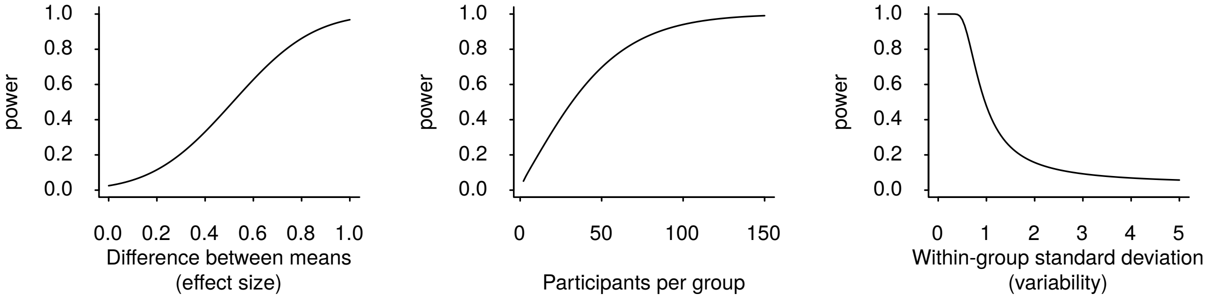


Figure 4.4: These three graphs show how the statistical power of a study varies with the effect size (left), the number of observations per group (middle) and the variability in the outcome variable within each group (right).

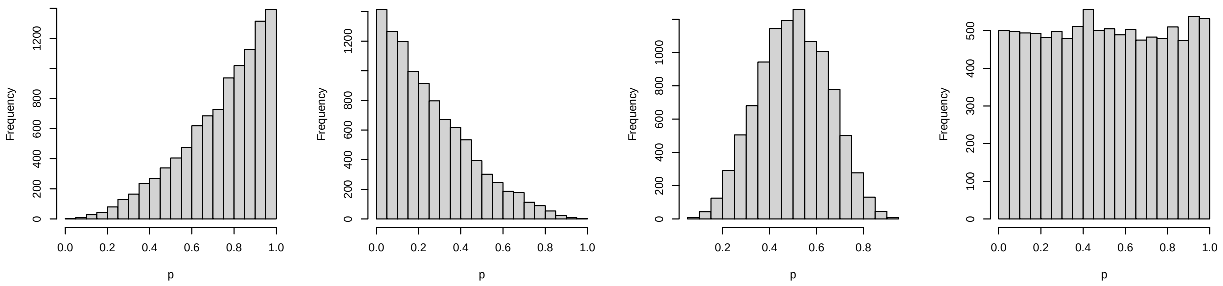


Figure 4.5: What could the distribution of 10,000 p-values look like under the null hypothesis?

5 Increasing precision

5.1 Precision

Up till now, our chief goal has been to increase the study's internal validity:

- Bias introduced by confounding can be countered by randomly assigning the participants (or whatever is being investigated) to the conditions. This won't always be possible, but randomisation remains the ideal.
- Bias introduced by expectancy effects, especially on the part of the researchers, can be reduced by blinding—for instance, by preventing raters from knowing which experimental condition the participant was assigned to.
- To decrease the chances that the results are affected by technical glitches or misunderstandings, the experiment should be piloted, and checks for comprehension and satisficing can be incorporated.

These three points concern **bias**—we want to prevent our study from systematically under- or overestimating the answer to the question we're interested in, which is often a causal question. But as we saw when discussing statistical tests, there is a random element to the results of any given study. In a study with random assignment, the luck of the draw may produce an estimated effect that is larger or smaller than the actual effect—it's just that randomisation helps to prevent this estimate from being *systematically* too large or too small. Roughly speaking, randomisation is equally like to yield overestimates as it is to yield underestimates.

But *an estimate obtained from an unbiased study can be completely off-target.* To appreciate this, consider a six-sided dice. The average number of pips on a six-sided dice is $\frac{1+2+3+4+5+6}{6} = 3.5$. Let's pretend we didn't know this and we wanted to estimate this number (i.e., 3.5) by throwing the dice and jotting down the number of pips showing face-up.

- If you do this just once, you'll just obtain an integer between 1 and 6 with equal probability (six possibilities). If you obtain a

1, your estimate will be off by 2.5 pips; if you obtain a 2, you'll be off by 1.5 pips; ...; if you obtain a 6, you'll be off by 2.5 pips. Taking into account all six possibilities, your average estimation error is 1.5 pips. Note that the estimation procedure itself is unbiased: underestimates and overestimates are equally likely to occur, and they're of the same size, so they will cancel each other out.

- If you throw the dice twice, you'll now observe one of $6^2 = 36$ possible outcomes. When you average the number of pips on both throws, you can still obtain an estimate of 1 (when you throw two 1s), but there's just a 1-in-36 probability of that happening. But 6 of the possible outcomes will be right on the mark (1+6, 6+1, 2+5, 5+2, 3+4, 4+3). Taking into account all 36 possibilities, your average estimation error is 0.97 pips. Again, this estimation procedure is unbiased.
- If you throw the dice five times, you'll observe one of $6^5 = 7776$ possible outcomes. When you average the number of pips on the five throws, there's just a 1-in-7776 probability that you'll end up estimating the average number of pips on the dice as 1. Taking into account all 7776 possibilities, your average estimation error is 0.62 pips. Again, this estimation procedure is unbiased.

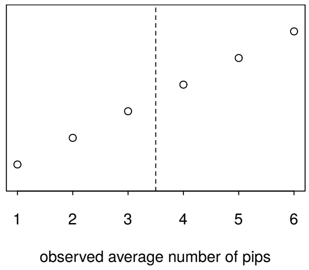


Figure 5.1: If you throw a six-sided dice once, you'll observe one of these 6 outcomes. The dashed vertical line highlights the true mean number of pips.

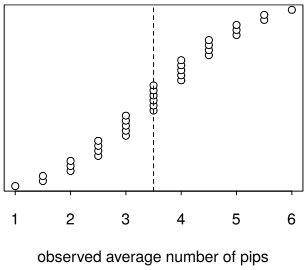


Figure 5.2: If you throw a six-sided dice twice and take the mean number of pips observed, you'll obtain one of these 36 outcomes.

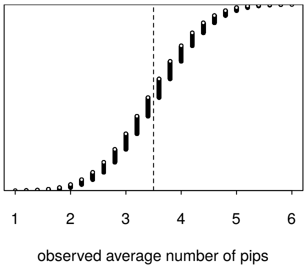


Figure 5.3: If you throw a six-sided dice five times and take the mean number of pips observed, you'll obtain one of these 7776 outcomes.

So as you increase the number of throws (i.e., as you increase the sample size), the average observation tends to correspond more closely to the true average. Put differently, your estimate tends to become more **precise**. It's still *possible* to be completely off mark, but it's less *probable*. Clearly, the third 'design' (throwing the dice 5 times) is preferable to the first and second design—not because it's unbiased (all three attempts are unbiased), but because the estimate it yields is expected to be closer to the truth.

In a similar vein, even unbiased studies can often be improved upon by taking steps that increase their precision. The precision of an estimate obtained in a study can itself be estimated and is typically expressed by means of **standard errors**, **confidence intervals**, or **credible intervals**. We won't concern ourselves here with how these statistics are to be calculated and interpreted; a rough appreciation of precision along the lines of the dice example suffices.

5.2 *Factors affecting precision*

Precision is affected mainly by the following two factors:

- the number of data points. Other things equal,[1] larger studies yield more precise estimates. As the dice example illustrates, the effect of increasing the sample size yields diminishing returns: the same number of added observations results in a greater increase in precision if the original sample is small compared to when it is large.

[1] This phrase is crucial. Large samples in and of themselves do not a good study make.

- the variability of the data within each group. The more variable the data within the groups are, the less precise the estimates will be. (In the dice example, the estimation error would be lower if our dice didn't have the values 1 and 6.) For the exercises in Section 4.6, you already identified a couple of ways to reduce the variability of the data within the groups (restricting the study to a more homogeneous group; using more precise measurements). But we can also reduce this variability through a combination of experimental design and statistics, see Sections 5.3 and 5.4.

Note that both of these factors also affect statistical power in the same way (see Figure 4.4 on page 38).

5.3 *Matching and blocking*

Matching Matching is a procedure in which researchers manually assign the participants (or whatever is being investigated) to the different conditions in such a way that both conditions are comparable on one or a number of background variables.

- Actual meaning: For each participant in condition A, find a similar participant (e.g., same age, sex and L2 skills) and assign this participant to condition B. This way, each participant has a counterpart in the other condition.
- What is often meant: Assign participants to conditions A and B in such a way that the *average* age (etc.) is similar or the same in both conditions. The individual participants themselves don't need to have any particular counterpart in the other group.

The rationale behind matching is that, by equating the conditions on one or a number of background variables, these variables can't act as confounding variables. However, matching is **not recommended**: It's possible that the researchers, while matching the participants on one background variable, inadvertently introduce a bias with respect to another background variable. Moreover, (pure) matching only allows you to equate those confounding variables that you matched for (see Figure 5.4). Randomisation also equates *other* (and indeed unknown) confounding variables and is superior to matching.

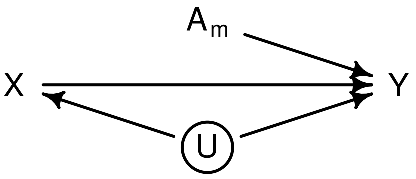


Figure 5.4: Matching the conditions (X) on A doesn't prevent confounding by other (perhaps unobserved) variables (U).

Blocking This is a type of matching, but unlike (pure) matching, it is used *in combination with* (rather than as an alternative to) randomisation.[2]

- Example 1: Based on a pool of participants, we build pairs of participants of the same sex and age and with a similar IQ. From each pair of participants, we *randomly* (rather than arbitrarily) assign one participant to condition A and one to condition B. In doing so, we both equate the two groups in terms of age, sex and IQ, but, due to the randomised assignment, we also prevent confounding by unobserved factors.[3]
- Example 2 (see Ludke et al., 2014): We randomly assign half of the female participants to condition A and half to condition B; same for the male participants. Again, randomised assignment helps to prevent confounding variables from biasing the results, and we have the added benefit that the two conditions will be perfectly balanced in terms of sex.

Blocking can increase a study's statistical precision—provided it is taken into account during the analysis.[4] The stronger the outcome is related to the blocking factors, the more powerful blocking is.

Note that blocking takes place *before* the random assignment. You can't block after the fact.

How does blocking increase precision? Let's say you're comparing the efficacy of two methods for learning Dutch as a foreign language. Six German-speaking and eight French-speaking learners sign up for your study, they work with one of the two learning methods for a while, and then take the same test at the end.[5]

- The speakers of German can be expected to have an advantage because of the similarity between Dutch and German. Let's say this advantage corresponds to 3 points on a 20-point test scale. (Realistically, you wouldn't be able to peg this number down so precisely.)
- Within each language group, learners still vary in their ability to learn Dutch.
- Let's say that, unbeknownst to you, learning method B yields a boost in test performance of 1 points relative to learning method A. Figure 5.5 shows what each learner's scores would have been like if they'd been tested on both methods.

There are $\binom{14}{7} = 3432$ ways to randomly assign these 14 participants to two conditions (A and B) with 7 participants each. For the learners in Condition A, we'd observe the scores shown as circles in Figure 5.5; for the learners in Condition B, we'd observe the scores shown as crosses. If we then took the mean difference between these groups, we'd end up observing one of the 3432 values

[2] "[M]atching is no real help when used to overcome initial group differences. This is not to rule out matching *as an adjunct to randomization*, as when one *gains statistical precision* by assigning students to matched pairs, and then randomly assigning one member of each pair to the experimental group, the other to the control group. In the statistical literature this is known as 'blocking."' (Campbell & Stanley, 1963, p. 15; my emphasis)

[3] This technique is extremely rarely used in our line of research, possibly because the pool of participants is rarely known at the start of the experiment.

[4] If you incorporate blocking in the design of your study, but you don't take this into account when analysing the data, you're not reaping its full benefits. For this class, you don't have to know how to take blocking into account in the analysis, just that you have to take it into account. But see Vanhove (2015) for some options.

[5] Again, the number of participants in this fictitious example is low to keep things tractable.

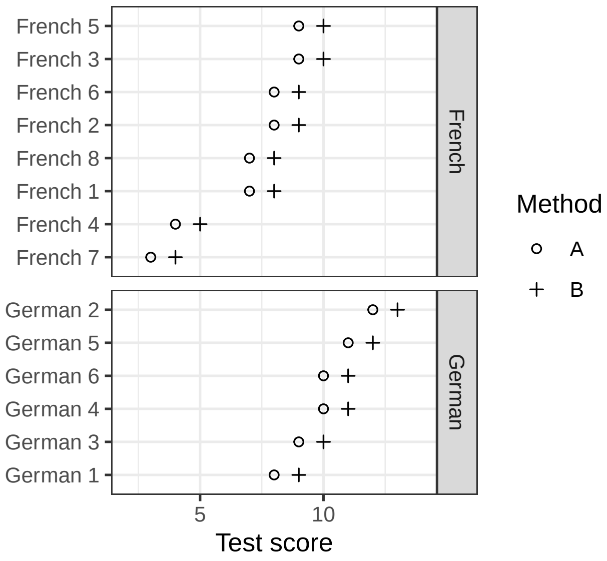


Figure 5.5: The test scores that each learner would obtain for each learning method. But we can't test all learners on one and the same method, so for half of the learners, we'll observe their scores for Method A, and for the other half, we'll observe their scores on Method B.

shown in Figure 5.6 on the next page. Complete random assignment would yield a mean estimation error of about 1.1 points.

However, we've identified a likely important source of variability in the data: language group. It makes sense to block on this factor, i.e., restrict the random assignment in such a way that half of the learners in each language group are assigned to one learning method and the other half to the other learning method. Of the 3432 total possible random allocations, only 1400 feature four French speakers in one condition and four in the other, as well as three German speakers in one condition and three in the other.[6] The average estimated difference between the two learning methods among these 1400 allocations is still 1, but the average estimation error is now only 0.9—an increase in efficiency of about 15%; see Figure 5.7. What has happened is that, by restricting the randomisation in this fashion, we limited both the number of allocations that would have yielded large overestimates (when the Method B condition would've consisted mainly of advantaged German speakers) and those that would have yielded large underestimates (when the Method B condition would've consisted mainly of disadvantaged French speakers). In our example, complete randomisation resulted in 476 out of 3432 (14%) allocations with an absolute estimation error of more than 2 points; blocked randomisation resulted in only 112 out of 1400 (8%) such allocations. Also see Table 5.1 on the following page.

So blocking on influential factors prevents the randomisation from generating some of the 'unlucky' allocations, thereby reducing the study's average estimation error, i.e., increasing its precision.

[6] There are $\binom{8}{4} = 70$ ways to split up the eight French speakers into two groups of four, and $\binom{6}{3} = 20$ ways of splitting up the German speakers into two groups of three. Combining these yields $70 \cdot 20 = 1400$ possibilities.

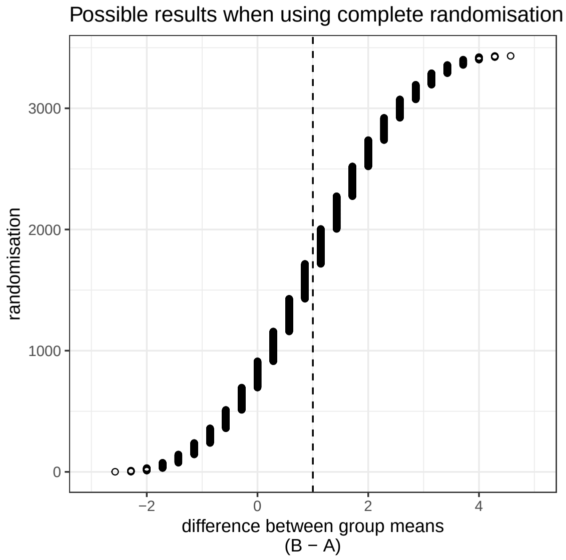


Figure 5.6: Across all 3432 possible random assignments, the mean estimation error of the difference between the two learning methods is about 1.1 points. The vertical line highlights the average estimate (viz., 1) across all 3432 possible random assignments.

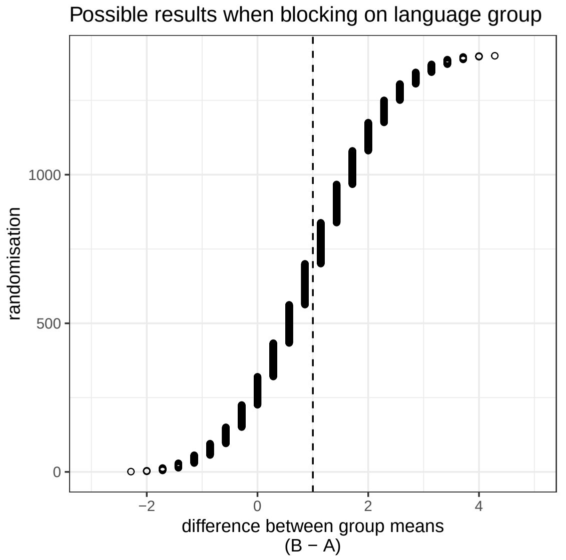


Figure 5.7: Of the 3432 possible random assignments, only 1400 have an equal number of French speakers assigned to each learning method, as well as an equal number of German speakers assigned to each learning method. Across these 1400 assignments, the mean estimation error is just 0.9. The average estimate (viz., 1), highlighted by the vertical line, doesn't change.

Interval	Complete randomisation	Blocked randomisation
(-3.5,-2.5]	0.03	0.00
(-2.5,-1.5]	2.16	0.93
(-1.5,-0.5]	12.70	9.79
(-0.5,0.5]	18.85	20.21
(0.5,1.5]	32.52	38.14
(1.5,2.5]	18.85	20.21
(2.5,3.5]	12.70	9.79
(3.5,4.5]	2.16	0.93
(4.5,5.5]	0.03	0.00

Table 5.1: In what percentage of randomisations did the mean difference end up in each interval? Note that for blocked randomisation, a greater percentage of randomisations yield a mean difference close to the true difference between the learning methods (i.e., 1) than for complete randomisation.

5.4 Leveraging control variables

Control variable Additionally collected variable that isn't of actual interest but that may account for differences between participants in terms of the outcome.

Example: the 'language aptitude test' in Ludke et al. (2014).

In randomised experiments, the added value of control variables is mostly statistical: If control variables can account for differences in the outcome between participants, they can be used to statistically reduce the variability within the groups. Similarly to blocking, this yields greater power and precision.[7] Note that the use of a control variable to this end is planned before the data are collected. Don't try out a bunch of 'control variables' during your analysis to see which works best!

Pretest Often, the most potent indicator of a participant's performance at the end of the experiment is their performance at the start of the experiment. A pre-intervention measure of their performance is therefore a useful control variable.[8]

It's also possible to 'block' on pretest scores. To this end, sort the participants according to their pretest score and divide them up into pairs like so: (12)(34)(56)(78).... Within each pair, randomly assign one participant to the control group and one to the intervention group. (You can similarly block on other continuous variables.)

Pre- and post-tests don't have to look identically. Any measure of pre-experiment performance is better than no measure at all.

Of course, if the pre- and posttests aren't similar and can't be scored on the same scale, you won't be able to make any claims about how much the participants progressed in each condition. *But that's not important!* What's important in a pretest/posttest design is the comparison between the conditions on the posttest scores. The purpose of the pretest scores is to increase the precision of this comparison. This can be achieved by using them as you would any other blocking or control variable, so they don't have to be expressed on the same scale as or be otherwise comparable with the posttest.

In fairly small studies, blocking tends to increase precision a bit more than merely using control variables in the analysis, but in the vast majority of cases, either is a good idea compared to the alternative of not leveraging any prior information![9]

Even if you're conducting a randomised experiment, it pays to think about which factors are likely to strongly affect the

[7] Contrary to common belief, including powerful control variables in the analysis is useful *even if* the groups are balanced with respect to these control variables. In fact, they're even more useful than when the groups aren't balanced.

[8] On taking into account pretest results, see Vanhove (2015).

[9] We don't need to concern ourselves with the freak cases where blocking or using control variables reduces precision (viz., tiny studies in which the blocking or control variables are uninformative with respect to the outcome; Imai et al., 2008.)

outcome so that, if feasible, you can take these factors into account using blocking or by means of control variables.

Don't go overboard with this, though. One or two strong blocking or control variables are likely to be helpful; umpteen variables that *might conceivably* bear some relation to the outcome aren't. Using several highly intercorrelated control variables isn't too useful either: they will all tend to do the same work, which makes them mutually superfluous.

Don't control for post-treatment variables! A fairly common error is that researchers control for variables that are themselves (directly or indirectly) affected by the treatment. The reason is that controlling for a 'descendant' of a variable is like controlling for the variable itself, only less strongly.

- If you 'controlled' for the treatment variable (e.g., throwing away data in order to keep it constant), you wouldn't be able to compare the outcome variable according to different values of the treatment variable (since there aren't any). Controlling for a descendant of the treatment variable (even by statistical means rather than by selecting observations) similarly amounts to throwing away data, just to a lesser extent. Rather than increasing power and precision, you'll lose some.

- If you 'controlled' for the outcome variable, you wouldn't be able to find any differences between the treatment groups even if the treatment produced some differences (since you fixed all outcome observations to the same value). Similarly, controlling for a descendant of the outcome variable (even by statistical means) typically amounts to artificially pulling the differences between the treatment groups towards zero.

See the DAGs in Figure 5.8 for these and two other cases.

Example 5.1. Say you want to find out if a pedagogical intervention boosts learners' conversational French skills. It may be a good idea to control for the learners' vocabulary knowledge. But if you collect the measure of vocabulary knowledge *after* the intervention, it's possible that this measure is also affected by the intervention. If you control for it, you could find yourself in one the situations depicted in Figure 5.8: learners could conceivably pick up some vocabulary as they're working on their conversational skills. ◇

A poorly chosen pretreatment control variable won't be too helpful, but it won't hurt your study either. But controlling for a posttreatment variable can bias your results or decrease their precision. Luckily, in true experiments, there's a simple solution:

Collect control variables at the outset of the study (before the intervention) so that you're sure that the control variables aren't themselves influenced by the intervention.

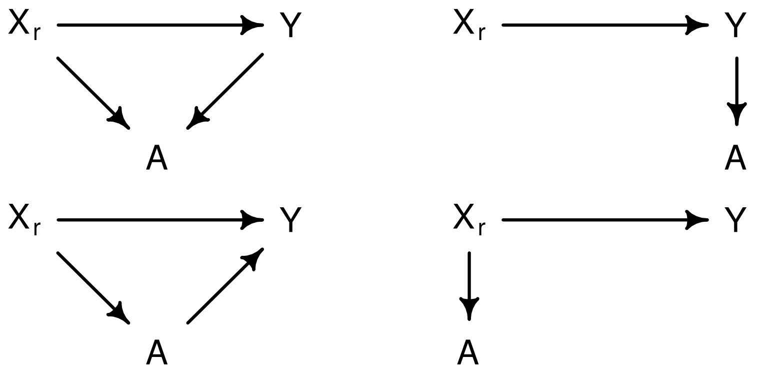


Figure 5.8: In all four DAGS, A is causally affected by X_r.

Top left: A is a collider. Controlling for it opens up a non-causal path between X_r and Y.

Bottom left: Controlling for A closes a causal path from X_r to Y. You'd be answering the question *What's the causal effect of* X_r *on* Y *other than that mediated by* A?. This may be interesting in its own right, but you have to be aware that this question differs from the question *What's the causal effect of* X_r *on* Y?

Top right: Controlling for A is like controlling for Y, just less strongly. If you want to estimate the causal effect of X_r on Y, controlling for Y would be a dreadful idea, so controlling for A would only be a slightly less dreadful idea.

Bottom right: Controlling for A isn't as terrible here as it is in the top right panel. But it's unnecessary, and you'll lose some precision.

Exercise 5.2. Consider the DAG in Figure 5.9.

(a) If the goal is to obtain an unbiased estimate of the causal influence of X on Y, which variable or which variables do you *need* to control for?

(a) Assume all variables shown in the DAG have already been collected (i.e., there are no further costs involved in collecting them). If the goal is to obtain an unbiased and maximally precise estimate of the causal influence of X on Y, which variable or which variables would you have to control for? ♢

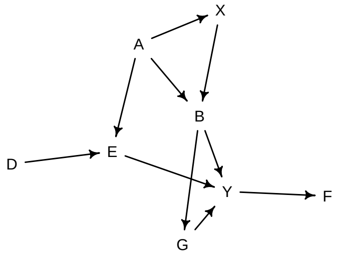


Figure 5.9: DAG for Exercise 5.2.

Exercise 5.3. Having taking this class, you've become the go-to expert on experimental methodology among your friends. So unavoidably, a friend of yours sollicts your input on an experiment she wants to run with one experimental condition and one control condition in a Zurich-based school class with 24 pupils. Of the fourteen pupils who have Swiss German as their main language, four receive special educational measures; of the ten pupils who do not have Swiss German as their main language, six receive special educational measures. Your friend also has the average school grade of each pupil.

After some discussion with your friend, you realise that, ideally, she should block on language background, whether the pupils receive special educational measures as well as average grade when randomly allocating the pupils to two equal-sized conditions.

(a) Explain to your friend what exactly she should do.

(a) How many different allocations can your randomisation scheme generate? Provide both the computation as well as the numeric result. (For comparison, complete randomisation could generate $\binom{24}{12} = 2704156$ different allocations.) ♢

Exercise 5.4 (Reading assignment)**.** The article by (Slavin et al., 2011) makes for pretty challenging reading, especially in terms of the analysis and the way the results are presented. But the introductory and methodological sections discuss some concepts that we've already discussed (especially p. 49) and introduces some new procedures.

First try to read the text in full, but skip the parts you find unintelligible. Then answer the following questions:

(a) "[C]hildren's reading proficiency in their native language is a strong predictor of their ultimate English reading performance." (p. 48, middle, left) What does this mean?

(b) What do the following terms mean?

  i. matching (p. 49, middle, right)

  ii. selection bias (p. 49, middle, right, and p. 50, top, left)

  iii. teacher/school effects (p. 49, bottom, right)

  iv. within-school design (p. 51, left). What would the opposite, a between-school design, look like?

(c) What is the independent variable in this study? What are the dependent variables?

(d) How were the pupils assigned to the different groups?

(e) Slavin et al. discuss at length how many pupils in each group (TBE vs SEI) couldn't be tested (Table 2) and whether the characteristics of these pupils differed between the conditions (Table 3). Why do you think they discuss this at all?

(f) "Children were pretested ... on the English Peabody Picture Vocabulary Test (PPVT) and its Spanish equivalent, the Test de Vocabulario en Imagenes Peabody (TVIP)." (p. 51, right) Why did the researchers go to the bother of conducting such pretests? Try to find at least two reasons.

(g) Comparing Slavin et al.'s Tables 4 and 6, we observe that the cohorts' average TVIP (*Test de Vocabulario en Imagenes Peabody*) scores dropped from 99.85 and 90.19 to 92.86 and 85.64, respectively, over the course of two years.

  Come up with the most plausible explanation for this result.

  Hint: The children's Spanish vocabulary knowledge probably did not decline on average as they got older. ◊

6
Pedagogical interventions

The following remarks are especially relevant for pedagogical interventions, but they apply to other studies, too.

6.1 Mortality

Less lurid and more descriptive terms are *drop-outs*, *outmovers*, and *panel attrition*.

In addition to lowering the sample size, drop-outs may bias the results of the study. As Figure 6.1 shows, being a drop-out or not can be a post-treatment factor, but one that has insidiously already been controlled for: the participants whose data goes into the analysis all have the same value on this variable.

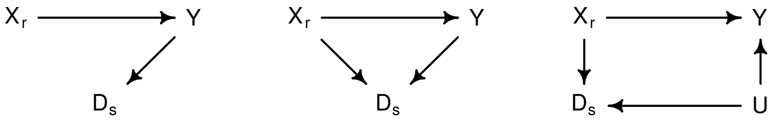


Figure 6.1: The fact that some participants stayed in the study and others can be treated as a factor in its own right (D_s; D for 'drop-out', $_s$ for making clear that selection took place).

Ideally, mortality doesn't depend on the condition (X), the participants' prior knowledge (U), or their progress (Y). When mortality does vary by condition, prior knowledge or learning progress, you have take into account the selection effect when interpreting the results.[1]

Figure 6.2 illustrates how mortality can bias the study's estimates. It is in fact possible that an observed 'treatment' effect is little more than a selection effect: If only gifted or highly motivated learners remain part of the treatment condition but the control condition isn't as selective, it's hardly surprising that at the end of the study, you'll find better scores in the treatment condition than in the control condition.

6.2 Clustering

In typical pedagogical settings, the participants can't be randomly assigned to the experiment's conditions on an individual basis. Instead, entire intact groups of participants (e.g., entire classes, entire schools, entire school districts etc.) are assigned to the same condition. This induces **clustering**: Due to teacher/class/school

[1] In a study with two measurement times, assessing learning progress for the drop-outs is impossible, but if you have several measurement times, you can check if those who progressed little thus far are more likely to drop out of the study.

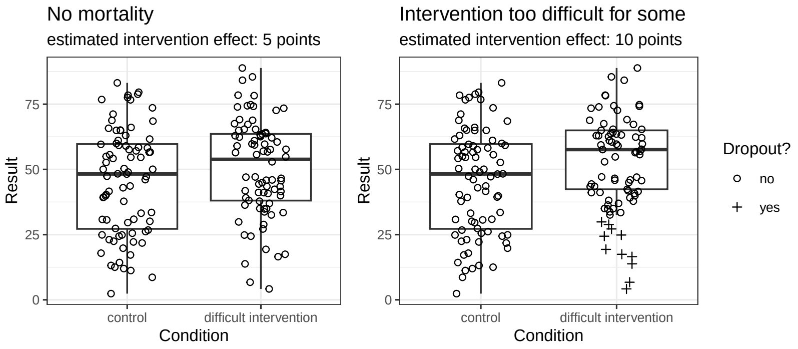
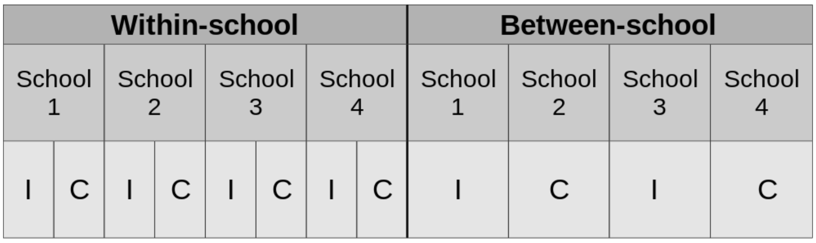


Figure 6.2: Example of the biasing effect of sample mortality. *Left:* Had it been possible to test all participants, we'd have found a mean difference of 5 points. *Right:* If the drop-out likelihood is itself affected by the intervention, we could end with a biased estimate.

etc. effects, participants belonging to the same cluster (class, school, etc.) tend to be somewhat more alike in their performance than participants belonging to different clusters.

Teacher/school effects

Within- vs between-school design See Figure 6.3. In a within-school (or similarly, within-class etc.) design, the school effect is neutralised using blocking, whereas the remaining possible confounding variables are ideally taken care of using randomisation. This increases precision relative to a between-school design. One possible drawback of a within-school design is that the pupils in the control and treatment classes in the same school influence each other (e.g., by comparing notes or helping each other make sense of what's being taught), which may wash out an existing treatment effect.



Within-school								Between-school			
School 1		School 2		School 3		School 4		School 1	School 2	School 3	School 4
I	C	I	C	I	C	I	C	I	C	I	C

Figure 6.3: Within- vs between-school designs. (I for intervention, C for control.)

Cluster-randomised design When clusters are assigned in their entiry to the same condition (e.g., all pupils in the same class are assigned to the same learning condition rather than each on an individual basis).

> In a cluster-randomised design, you need to take the cluster-randomisation into account when analysing the data.

To appreciate the need for taking clustering into account during the analysis, consider an experiment with 6 classes of 10 pupils each. There are over one-hundred *quadrillion* ways to split up 60 pupils into two groups of 30 ($\binom{60}{30} \approx 1.2 \times 10^{17}$). But there are only 20 ways to split up six classes into two groups of three classes each. The analysis needs to be based on the assumption that the allocation obtained is one of 20 possible ones, not one of a gazillion ones.

You need to be able to recognise a cluster-randomised design and to know that you need to take the cluster-randomisation into account during the analysis. But for this class, you don't need to know how to take it into account.[2] If you want to conduct an experiment that uses cluster-randomisation, see Vanhove (2015), Vanhove (2020), and references therein.

What you also need to know about cluster-randomisation is this:

> Studies with one intact group as the experimental group and another intact group as the control group are useless.

The reason is that class, school and teacher effects can't be separated from the effect of the intervention if you just have one intact group per condition.

Exercise 6.1. Let's say you want to run a pedagogical experiment (e.g., to compare two learning methods for French as a foreign language) in which randomisation has to take place at the class level rather than at the individual level. Other things equal (e.g., number of classes, number of pupils), what are the advantages and drawbacks of the following designs? What's the worst option? What's the best?

(a) All classes are taught by the same teacher.

(b) Each class is taught by a different teacher.

(c) One teacher teaches all classes in the control condition, and another teacher teaches all classes in the intervention condition.

(d) Each teacher teaches two classes: one in the control condition, and one in the intervention condition. ◇

Exercise 6.2 (Reading assignment)**.** The study by Kang et al. (2013) serves as an example of a research design we haven't encountered yet. Additionally, it uses some turns of phrase commonly found in research reports:

(a) Are both studies experiments with control groups?

(b) "The Hebrew nouns were learned in one of two training conditions – retrieval practice or imitation – that were manipulated within subjects across separate blocks and semantic categories." (p. 1261)

   i. What does "manipulated within subjects across separate blocks" mean?

[2] One simple but valid approach is to compute the mean of each cluster and then analyse these means instead of the raw data.

   ii. What does "manipulated within subjects across semantic categories" mean?

(c) p. 1262:

   i. "The order of items in each test was randomized for each learner." Why?

   ii. "In Experiment 2, the order of both tests was counterbalanced across learners. . . " "Counterbalanced across learners" is experimental jargon. It merely means that half of the learners first took Test A and then Test B, whereas the other half first took Test B and then Test A. But why didn't all learners simply take the tests in the same order?

(d) "The α level for all analyses was set at .05" merely means that p-values smaller than 0.05 were regarded as statistically significant. This is rarely mentioned explicitly. ♢

7 Within-subjects experiments

7.1 Advantages and drawbacks

Blocking increases power and precision by pairing up similar participants and randomly assigning one of each pair to each condition. In within-subjects designs, this idea is taken to an extreme: the *same* participants are tested in the different conditions.

> In a within-subjects experiment, every participant serves as their own control.

Advantage 1: Easier to explore interindividual differences With a between-subjects experiment, you can only estimate the average effect of an intervention. Within a within-subjects experiment, you can additionally gauge which participants gain more from an intervention than others.

Advantage 2: Statistical precision A study's statistical precision depends on (a) the amount of data and (b) the variability in the data. The *main* (!) advantage of a within-subjects design is that it easily accounts for an important source of variability: interindividual differences.

How much more precise a within-subjects experiment is than a between-subjects experiment varies from case to case.[1]

Possible drawback 1: Lack of ecological validity (Not too relevant for basic research.) In applied settings, you typically want the study to mimic the context in which its findings are to be implemented. But in such a context, people (e.g., pupils) won't be exposed to several conditions (e.g., learning methods) but rather to just one.

Possible drawback 2: Order and carry-over effects When participants are tested in several conditions, it's possible that they learn something in one condition that affects their performance in the other condition (**carry-over effect**). It's also possible that their performance in the last condition differs from their performance in the first condition because they've had more practice or because they've grown tired of being tested (**order effects**). Luckily, these dangers can be minimised.

[1] Quené (2010) estimates that within-subjects designs have the statistical precision of between-subjects designs with four times as many participants. The precise factor depends on the extent to which the participants' performance in one condition correlates with their performance in the other condition: The stronger this correlation, the greater the added value of a within-subjects experiment. But even if you can't quantify this added value: Within-subjects designs offer more statistical precision.

7.2 Minimising order effects

Complete counterbalancing To prevent learning or fatigue effects from exerting a systematic effect on the results, you can vary the order of the conditions between the participants. In complete counterbalancing, *all* possible orders are taken into account. If you have two within-subjects conditions, half of the participants first complete condition A and then B, and the other half first complete condition B and then A. If you have three conditions, there are six possible orders[2]; you can then randomly assign one sixth of the participants to first complete A, then B, then C; one sixth completes A, then C, then B, etc.:

A	A	B	B	C	C
B	C	A	C	A	B
C	B	C	A	B	A

Table 7.1: Complete counterbalancing for a within-subjects experiment with three conditions (or stimulus sets and the like).

Latin squares If a within-subjects experiment has lots of conditions, complete counterbalancing is impractical. For four conditions, for instance, there are already 24 possible orders—we may not even have that many participants! The Latin square lends itself to such cases.[3] Latin squares are arrangements of symbols in a grid in which each of the symbols used occurs exactly once in each row and exactly once in each column. The grid below is a Latin square of size 4—one of the 576 possible arrangements of the symbols A, B, C, and D that form a Latin square. (Can you come up with a couple of the other 575 ones?)

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

Table 7.2: A Latin square for a within-subjects experiment with four conditions (or stimulus sets and the like).

Let's say you picked the Latin square above for your study. You'd then randomly assign one quarter of the participants to the condition (or stimulus set) order ABCD (first row), one quarter to BCDA (second row), one quarter to CDAB (third row) and one quarter to DABC (fourth row). The conditions (or stimulus sets) are randomly assigned to one of the letters, too.

Other possibilities In which order should we show our participants 50 pictures that they are to describe if we want to prevent order effects from biasing the results?[4] $50! = 3 \cdot 10^{64}$ is an astronomical number, and even just 50 different Latin square orders seem impractical. One of several possible solutions is to present the images in a new random order for each participant. The drawback of doing this is that perhaps image 3 occurs much more often at the start than at the end of the data collection.

In many psycholinguistic studies, participants need to react to several stimuli per condition (e.g., 12 stimuli per condition). The

[2] $3! = 3 \cdot 2 \cdot 1 = 6$.

[3] 'Latin' because the symbols used are typically letters of the Latin alphabet.

[4] Of course, it's possible that we just accept such 'bias' if we aren't interested in differences between the images, but just in differences between the participants. If that's the case, these steps may be superfluous.

order of the stimuli in these studies are often randomised so that the conditions are mixed up (e.g., ABAABBBAB etc.).

If you have a genuine choice between a between-subjects and a within-subjects design for your own research, pick the within-subjects design. (Unless, of course, you have an excellent reason not to do so.)

The possible danger of carry-over effects typically isn't large enough to offset the certain gain in statistical precision.

Exercise 7.1. From Ludke et al. (2014):

> Participants were randomly assigned to one of three learning conditions: speaking, rhythmic speaking, and singing. The participants heard 20 paired-associate phrases in English and an unfamiliar language (Hungarian) (...). (...) The 15-min learning period was followed by a series of five different production, recall, recognition, and vocabulary tests for the English–Hungarian pairs.

Re-design this between-subjects experiment as a within-subjects experiment. What would this description look like? For the time being, ignore the *rhythmic speaking* condition.

In your design, you are constrained by the resources Ludke et al. had. This means that you cannot introduce stimuli and tests that were not already used by Ludke et al., and that you have to work with at most 30 men and 30 women as participants.

(There are several acceptable solutions.) ♢

Exercise 7.2. As in the previous exercise, but with all three conditions. (There are several acceptable solutions.) ♢

7.3 *Optional: Further reading*

Latin-square designs are used, albeit less commonly, in studies other than within-subject experiments. See Richardson (2018) for an overview and some finer points that weren't discussed here; his article is geared towards educational researchers.

Exercise 7.3 (Reading assignment)**.** The Simon task was (and still is) commonly used in research on any cognitive advantages of bilingualism. In a nutshell, the theory is that bilinguals have to constantly inhibit one of their languages. Because of this, they practice their 'inhibitory control'. The Simon task is purported to tap into this same skill (suppressing impulses). As a result, some researchers have interpreted smaller Simon effects in bilinguals as evidence for cognitive advantages of bilingualism. However, these studies have drawn criticism (see the references in Kirk et al., 2014).

Read pages 640–643 of Kirk et al. (2014). Don't read the Results section (it is needlessly complicated), but do take a look at Table 1 and Figure 1. Then answer the following questions; short answers suffice.

1. Quasi-experiments tend to be both less conclusive and more arduous than true experiments. Explain why.

2. "Colour assignment to key location was counter-balanced across participants." (p. 643, top right) Rewrite this sentence without using the word *counter-balanced*. Do you have any idea why the researchers effected this counter-balancing? If so, briefly describe it.

3. On page 641, the authors make two predictions:

   - "If bilingualism, rather than ethnic or cultural background is linked to superior executive control, then both bilingual groups should exhibit an advantage compared to the monolinguals."
   - "If use of multiple dialects incurs executive control benefits similar to those observed in bilinguals, then differences in executive control between bidialectal monolinguals and bilinguals might be attenuated."

   Do the results in Figure 1 suggest that "both bilingual groups . . . exhibit an advantage compared to the monolinguals"? Do they suggest that any differences in executive control between bidialectal monolinguals and bilinguals is smaller than those between non-bidialectal monolinguals and bilinguals? Explain precisely which comparisons you base your answers on.

4. Assume you had access to Kirk et al.'s full data set and that you were asked to draw a more informative graph than Figure 1. Draw a sketch of what your graph would look like and briefly explain why your graph is better than Kirk et al.'s. (You do *not* need to draw such a graph in R. A free-hand sketch suffices. Do make sure, however, to indicate clearly what it is that you want to plot. Label your axes.) ◊

8
Quasi-experiments and correlational studies

Whether a study counts as a quasi-experiment or a correlational study depends on whom you ask. Some researchers use the term *quasi-experiment* to refer to cluster-randomised experiments, whereas others use the term *pre-experiment* to refer to group comparisons without randomisation. Similarly, different researchers draw the border between quasi-experiments and correlational studies at different places. As far as I'm concerned, the communalities between the two outnumber the differences. For what it's worth, I use *quasi-experiment* for group comparisons and *correlational study* when the predictor is continuous. What's important is that, because no random assignment was used, we can't assume that the treatment variable is independent of pre-treatment variables—confounding is a real threat.

Quasi-experiment Group comparison, but the groups weren't constructed using random assignment.

- Example 1: Comparison of pupils with and without an immigration background.
- Example 2: Comparison of kids that take heritage language classes and kids that don't.

It doesn't matter whether the groups *could* have been constructed using random assignment, just whether they were.

Correlational study No group comparison. Instead, one assesses to what extent variation in an outcome (dependent) variable can be accounted for by differences in one or more continuous predictors (independent variables). The values of these predictors weren't assigned to the units of observation randomly. More often than not, control variables are taken into account as well.

- Example 1: Researchers collect IQ and L2 proficiency data in a group of learners and assess how strongly both types of data covary.
- Example 2: Using archival data, researchers gauge how well they can account for whether children will pass their A-levels based on the results of a vocabulary test when the children were 12 years old.

Why carry out quasi-experiments and correlational studies?

> "But just because full experimental control *is* lacking, it becomes imperative that the researcher **be thoroughly aware of which specific variables his particular design fails to control**.
>
> "The average student or potential researcher reading the previous section of this chapter probably ends up with more things to worry about in designing an experiment that he had in mind to begin with. This is all to the good if it leads to the **design and execution of better experiments and to more circumspection in drawing inferences from the results**. It is, however, an unwanted side effect if it creates a feeling of hopelessness with regard to achieving experimental control and leads to the abandonment of such efforts in favor of even more informal methods of investigation.
>
> "[W]e shall . . . survey the strenghts and weaknesses of a heterogeneous collection of quasi-experimental designs, each deemed worthy of use *where better designs are not feasible*." (Campbell & Stanley, 1963, p. 34; their emphasis in italics, mine in bold-face)

Note that the goal of quasi-experiments and correlational studies is often to draw causal conclusions, but the findings—for better or for worse—tend to be couched in non-causal language (Grosz et al., 2020).

Controlling for confounds is difficult Consider the following description.

> "There were 40 participants who composed two language groups and two age groups. Twenty of the participants were younger adults ranging in age from 30 to 54 years (mean age = 43.0 years), and 20 were older adults ranging in age from 60 to 88 years (mean age = 71.9 years). In each age group, half the participants were monolingual English speakers living in Canada, and the other half were Tamil–English bilinguals living in India. (. . .) All the participants in both groups had bachelor's degrees . . . " (Bialystok et al., 2004, p. 44)

While the authors didn't explicitly claim to have done so, you might end up thinking that level of education was controlled for in this study. A couple of minutes' thought should reveal that this wasn't the case. (Does having a minimum requirement of having Bachelor's degrees equate both groups with respect to level of education?) But more interestingly, by introducing this minimum requirement, the authors may have introduced *additional* bias. How so?[1]

8.1 Correlation coefficients

Pearson correlation coefficients, typically just called correlation coefficients and abbreviated as r, express how closely the (X,Y) data points fall on a straight line.

- r = 1: All points fall exactly on an increasing line.
- r = −1: All points fall exactly on a decreasing line. Correlation coefficients of 1 or −1 (or close to it, e.g., r = 0.99) tend not to

[1] If you're stuck, consult `https://gpseducation.oecd.org/CountryProfile?primaryCountry=CAN&treshold=5&topic=EO` and look up similar data for India.

be too interesting: They typically indicate that the two variables express the same thing (e.g., body length in centimetres and in inches).

- r = 0: There's no linear relation between the two variables whatsoever.

Correlation coefficients work in both directions: $r_{XY} = r_{YX}$.

Figure 8.1 shows eight examples of scatterplots and the correlation coefficients for the data presented in them. Note that a correlation coefficient close to zero doesn't imply that there is no relation between them; correlation coefficients different from 1 or -1 don't imply that the relation between two variables is imperfect; and it's possible for a positive correlation coefficient to reflect a relationship that's largely negative, and vice versa. Do these examples contradict the rough definition of correlation coefficients given above?

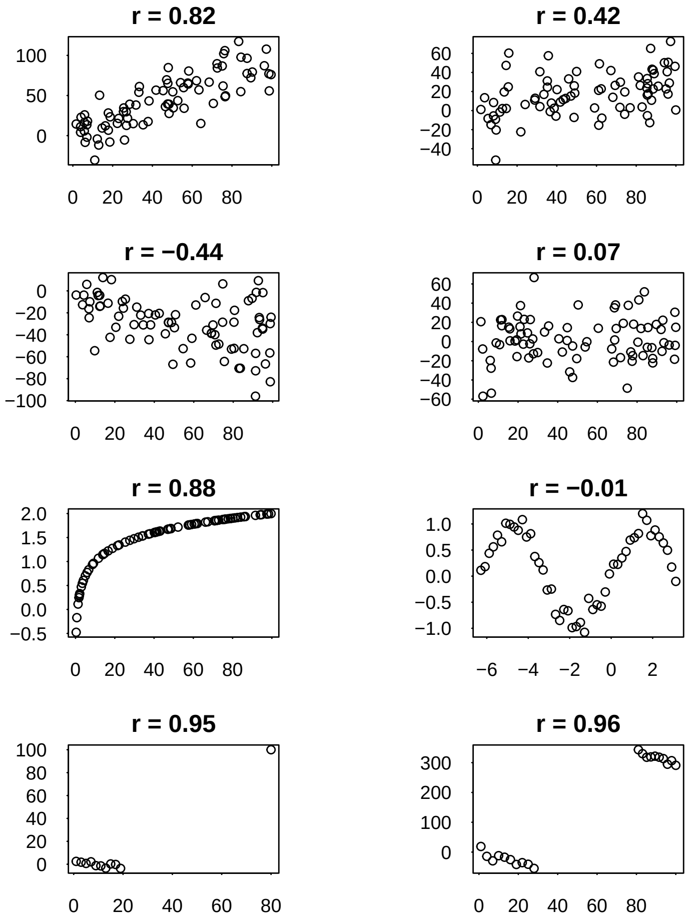


Figure 8.1: Examples of scatterplots and their associated correlation coefficients.

The same correlation coefficient can correspond to a multitude of relationships between two variables. Never *ever* compute a correlation coefficient without drawing a scatterplot first.

Honing your intuitions about correlation coefficients To hone your intuitions about correlation coefficients, you can use the `plot_r()` function from the `cannonball` package for R.[2]

```
# Install the package
install.packages("devtools")
devtools::install_github("janhove/cannonball")

# Load the functions
library(cannonball)

# Draw 16 plots with 20 data points each and r = 0.6
plot_r(n = 20, r = 0.6)

# With 50 data points each and r = 0.0
plot_r(n = 50, r = 0.0)

# With 40 data points and r = -0.9
plot_r(n = 40, r = -0.9)
```

Type `?plot_r` at the R prompt to access the function's help page and read the text under 'Details'.

8.2 Statistical control using hierarchical regression

In correlational studies, control variables are often used to adjust statistically for known confounders. One technique used to accomplish this is hierarchical regression; see Table 3 in Slevc & Miyake (2006) for an example. We will discuss this technique mainly so that you are better able to appreciate the shortcomings of this technique and ones similar to it.

Example If you measure the shoe size and vocabulary knowledge of 4- to 16-year-olds, you'll observe a positive correlation between the two. This isn't surprising; see Figure 8.2.

We'll use this silly example to illustrate the principle behind hierarchical regression; see Figure 8.3.

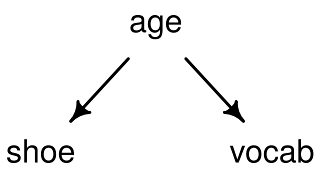


Figure 8.2: Shoe size and vocabulary knowledge are correlated since age acts as a confound.

- Top left: Shoe size and vocabulary knowledge are positively correlated.
- Top right and middle left: Age—the confound—is correlated positively with both shoe size and vocabulary knowledge.
- Middle right: This plot shows the vertical distance between the points in the middle left panel and the regression line. This shows how much the participants vary in their vocabulary test scores once the linear association between age and vocabulary knowledge has been partialled out.

[2] If you run into the error message 'Failed to install 'unknown package' from GitHub', try running the command `Sys.unsetenv("GITHUB_PAT")` first.

- Bottom left: The association between shoe size and the vocabulary test scores with the linear association of age partialled out is much less strong. In this simulated example, the fact that the remaining association isn't exactly zero is due entirely to chance.

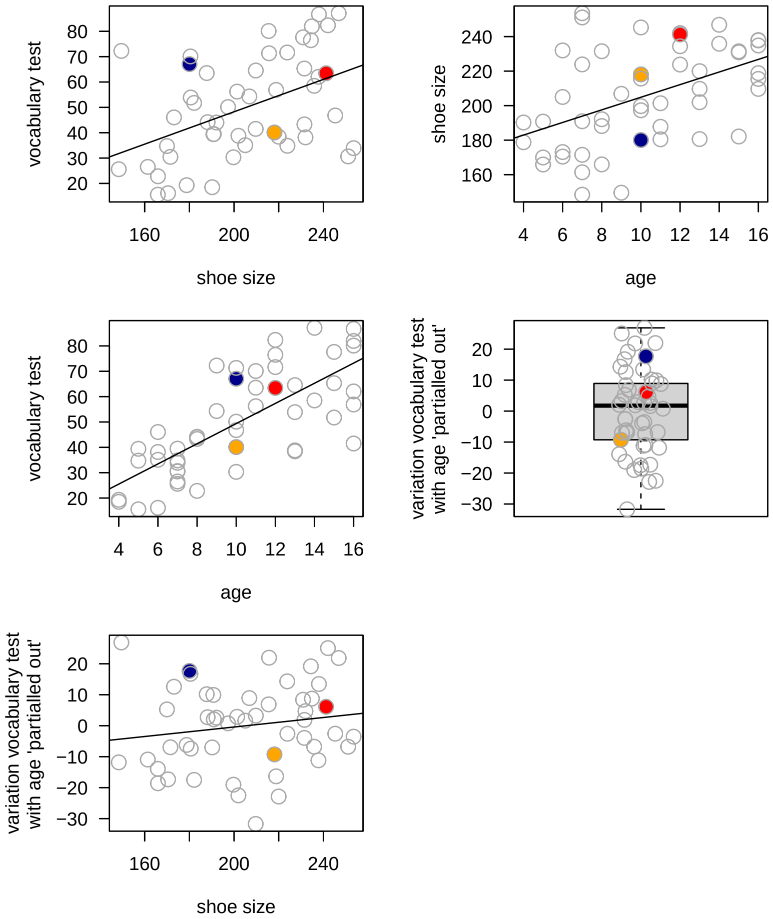


Figure 8.3: Hierarchical regression used to control for the age confound in the relationship between shoe size and vocabulary knowledge. The coloured circles in each panel show data belonging to the same three participants. The straight lines are **regression lines**. This is the straight line that best captures the tendency in the cloud of data points.

8.3 *Caveats*

You need to be hyper-aware of the following caveats concerning statistical control:

1. Controlling for a number of possible confounds doesn't rule out the possibility that there are even more confounds; Figure 8.4.

2. The methods typically used to account for confounding variables account for *linear* relationships between the confounds and the variables of interest. If these relationships aren't linear, the confound won't be fully accounted for. In DAG parlance, the path via the confound won't be fully closed.

3. The 'confound' may be a post-treatment variable. See Section 5.4 on page 45.

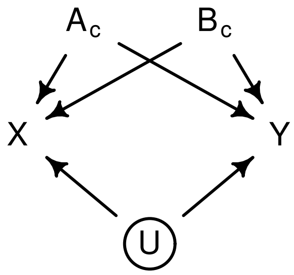


Figure 8.4: Perfectly controlling for A and B closes the non-causal paths X ← A → Y and X ← B → Y. But it leaves open the non-causal path via U.

4. Statistical control may be imperfect because the confound was measured with some error. We'll treat this in more detail in Chapter 9.

   The following excerpt makes the same points:

> "When experimental designs are premature, impractical, or impossible, researchers must rely on statistical methods to adjust for potentially confounding effects. Such procedures, however, are quite fallible. We examine several errors that often follow the use of statistical adjustment.
>
> "The first is inferring a factor is causal because it predicts an outcome even after "statistical control" for other factors. This inference is fallacious when (as usual) such control involves removing the linear contribution of imperfectly measured variables, or when some confounders remain unmeasured.
>
> "The converse fallacy is inferring a factor is not causally important because its association with the outcome is attenuated or eliminated by the inclusion of covariates in the adjustment process. This attenuation may only reflect that the covariates treated as confounders are actually mediators (intermediates) and critical to the causal chain from the study factor to the study outcome.[3]
>
> "Other problems arise due to mismeasurement of the study factor or outcome, or because these study variables are only proxies for underlying constructs.
>
> "*Statistical adjustment serves a useful function, but it cannot transform observational studies into natural experiments, and involves far more subjective judgment than many users realize.*" (Christenfeld et al., 2004, abstract, my emphasis)

Large sample sizes don't solve these problems.

Also see the blog entry *Controlling for confounding variables in correlational research: Four caveats*.

Exercise 8.1. In this series of exercises, we will use R to simulate some data and perform correlation analyses on them. More concretely, we will generate $n = 200$ pairs of predictor–outcome observations. The predictor $x = (x_1, \dots, x_n)$ will be sampled from a normal distribution with mean 0 and variance 1 (i.e., $x \sim \mathcal{N}(0, 1)$). The outcome y is described by a simple function of x and some random error, namely,

$$y_i = 0.4 + 0.7 \cdot x_i + \varepsilon_i,$$

$i = 1, \dots, n$, where the random error $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ is also sampled from a normal distribution with mean 0 and variance 1 (i.e., $\varepsilon \sim \mathcal{N}(0, 1)$).

Run the following commands in R to simulate data according to this scheme. Note that the `rnorm()` command takes an argument that specifies the standard deviation (`sd`) of the normal distribution from which to sample the data, not its variance. But since $\sqrt{1} = 1$, the standard deviation of the distribution from which we sample

[3] What's meant is a causal chain such as A → B → C. A is causally important, but if you control for B, you won't find any association between A and C.

is also 1. (Like for the graphing assignments, don't enter these commands directly to the console. Use a script.)

```r
n <- 200
x <- rnorm(n, mean = 0, sd = 1)
y <- 0.4 + 0.7 * x + rnorm(n, mean = 0, sd = 1)
```

The expected correlation between x and y is

$$\rho_{xy} = \frac{0.7\mathrm{Var}(x)}{\sqrt{0.7^2\mathrm{Var}(x) + \mathrm{Var}(\varepsilon)}\sqrt{\mathrm{Var}(x)}} = \frac{0.7}{\sqrt{0.7^2 + 1}} \approx 0.57,$$

since $\mathrm{Var}(x) = 1 = \mathrm{Var}(\varepsilon)$.

We now put these x, y observations into a tibble.

```r
library(tidyverse)
d <- tibble(predictor = x, outcome = y)
```

1. Using `ggplot()`, draw a scatterplot of the `predictor` vs. `outcome` values in `d`. (You don't have to hand this in.)

2. Compute the sample correlation coefficient between `predictor` and `outcome` using the R function `cor()` like so:

   ```r
   cor(d$outcome, d$predictor)
   ```

   Jot down this number, rounded to two decimal places. Now simulate the data again and jot down the number again; do this five times. Compare the correlation coefficients you observed to the expected correlation coefficient and summarise your findings.

   Hand in the five sample correlation coefficients you recorded, rounded to two decimal places, as well as a succinct summary.

3. Now imagine that instead of observing a random sample of n observations, we only observe data from those pairs of observations where $x > 0$. To emulate this scenario, we create a new tibble (`d2`) consisting of those rows in `d` where the predictor value is greater than 0:

   ```r
   d2 <- d |>
     filter(predictor > 0)
   ```

   Draw a scatterplot of the `predictor` vs. `outcome` values in `d2`. (You don't have to hand this in.) Compute the sample correlation coefficient between `predictor` and `outcome` in this reduced sample. Repeat the entire simulation five times. Compare the correlation coefficients you observed to the expected correlation coefficient and summarise your findings.

   Hand in the five sample correlation coefficients you recorded, rounded to two decimal places, as well as a succinct summary.

4. Now imagine that instead of observing a random sample of n observations, we only observe data from those pairs of observations where $x > 1$ or where $x < -1$; equivalently, where $|x| > 1$. That is, we're only retaining data with fairly extreme x observations.

   Using `filter()`, create a new tibble (`d3`) consisting of those rows in `d` where the absolute predictor value is greater than 0. You can use the `abs()` function to compute absolute values.

   Draw a scatterplot of the `predictor` vs. `outcome` values in `d3`. (You don't have to hand this in.) Compute the sample correlation coefficient between `predictor` and `outcome` in this reduced sample. Repeat the entire simulation five times. Compare the correlation coefficients you observed to the expected correlation coefficient and summarise your findings.

   Hand in the five sample correlation coefficients you recorded, rounded to two decimal places, as well as a succinct summary.

◊

9
Constructs and indicators

We're faced with an inescapable fact of life.

> Most measurements are imperfect.

Saying that a study's measurements aren't perfect isn't much of a criticism. But it's crucial to appreciate the consequences of imperfect measures—pointing out that a study's findings can plausibly be accounted for by the fact that its measurements are imperfect *is* a valid criticism.

Construct or *latent variable*. Lots of characteristics can't be observed or measured directly. Instead, their existence, as well as their relative value, are inferred on the basis of other, observable variables.

Indicator or *manifest variable*. These are variables that can be measured or observed directly and from which information about the construct is inferred. Table 9.1 lists some examples.

Table 9.1: Examples of constructs and indicators.

Construct	Example indicator
Intelligence	Your result on an intelligence test
Working memory capacity	The length of a sequence of digits you can repeat in reversed order
Language aptitude	Your result on the LLAMA-D test
L2 reading skills	The number of correctly answered items on a reading test
Attitudes towards Danish	Your answer to the question 'How beautiful do you think Danish is?'
Socio-economic status	Your father's occupational category

Measurement error Even the best indicators are rarely perfect. Better indicators just have a smaller measurement error.

Even variables that don't act as a proxy for some cognitive or social construct are often measured with some error. Examples include body weight (bathroom scales are imperfect, and the result is rounded), blood pressure (if you have a sphygmomanometer[1], check its manual), and age (invariably rounded down to the integer below when reported).

9.1 *Systematic and random measurement error*

Measurement error can include both a systematic and a random component.

The **systematic** component of an instrument's measurement error is the extent to which it tends to over- or underestimate what it's supposed to measure. For instance, a miscalibrated kitchen scale may overestimate weights by 10 g on average, and an overly harsh language test may tend to label learners' L2 skills one CEFR level below their actual proficiency on average.

Note that it's possible for an instrument to systematically overestimate values on one part of the scale and to underestimate them on another part.

When there's no gold standard to which the measurements can be compared, it may be impossible to assess their systematic measurement error.

The **random** component of an instrument's measurement error is the extent to which the measured values differ from the true values + systematic error. Another way of putting this is: By how much will the measurements vary if the true values are the same? For instance, an kitchen scale may, on average, measure weights accurately (no systematic error), but the individual readings may be off by up to a couple of grams in either direction (random error).

As a second example, consider a group of 365 7-year-olds, all born on different days of the year. Just one of them actually is 7 years old on the day; the reported values of the others will be off by 1 day, 2 days, ..., 364 days. The reported age, then, systematically underestimates their true age by $\frac{0+1+2+\ldots+364}{365} = 182$ days. The random component is 0, though, as children born on the same day will report the same age, even though this reported age will be lower than their actual age.

As a final example, consider a poorly calibrated bathroom scale. If you put a calibrated mass of precisely 60 kg on it on five different occassions, it returns readings of 61.1, 60.4, 60.4, 60.5 and 61.2. The mean observation for the same mass is $\frac{61.1+60.4+60.4+60.5+61.2}{5} = 60.78$, i.e., an overestimate of 0.78 kg. The mean absolute difference between the observations and their expected value (here: 60.78) is $\frac{|0.32|+|-0.38|+|-0.38|+|-0.28|+|0.72|}{5} = 0.42$.[2]

[1] I had to look this up.

[2] You won't have to do such calculations yourself. The main thing is that you appreciate the difference between the systematic and random components of measurement error.

9.2 *Consequences of measurement error*

The consequence of systematic measurement error is clear: Your data are biased. This isn't necessarily a problem: If you're comparing two groups for both of which you have data that are biased to the same extent, the difference between them won't be biased. And for variables such as age, the systematic error (roughly 182 days) tends to be small relative to the variability of the true values, in which case it's probably inconsequential.[3]

The consequences of random measurement error are much less intuitive and bear pointing out.

Less power and precision Measurement error on the *outcome* variable will increase its variability. Since power and precision are lower when there's more variability in the outcome, measurement error on the outcome lowers power and precision.

Statistical control is imperfect Measurement error on a *control* variable means that controlling for this observed variable won't fully eradicate the confounding caused by the construct itself. The DAG in Figure 9.1 illustrates this.

> "[F]allibility in a covariate usually implies that there would be more adjustment if the variable were measured without error." (Huitema, 2011, p. 569)

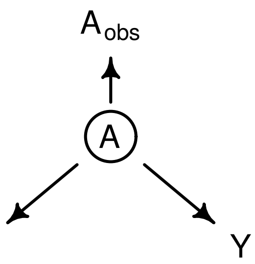


Figure 9.1: The X–Y relationship is confounded by A. A, however, can't be observed directly. A proxy (indicator) A_{obs} can be controlled for instead, but this won't fully shut the non-causal path X ← A → Y.

Controlling for A_{obs} is better than not controlling for it. But researchers routinely mistake controlling for an indicator with controlling for a construct, and their causal conclusions are overconfident as a result. A discussion of this problem can be found in Westfall & Yarkoni (2016), Vanhove & Berthele (2017) and Berthele & Vanhove (2020).

Regression to the mean When observations are due partly to skill or some underlying construct and partly to chance (e.g., measurement error), a second round of observations will likely show that the extreme scores have become less extreme, i.e., they've regressed to the mean.

- First consider an example where the observations are purely due to luck, with no skill or construct involved: playing roulette. Playing roulette is a losing proposition: For every 100 francs bet, you stand to lose about 5 francs (= the mean). But on any given night, some players will luck out and make a killing, whereas other players get extraordinarily unlucky and lose much more than the expected 5 francs. Their winnings or losses are a dreadful measure of their skill level: they all have the same skill level, which corresponds to a loss of 5 francs.

  The next night, however, the lucky players from the day before probably won't get as lucky again (their luck the day before was

[3] But see, for instance, Helsen et al. (2005) and Sprietsma (2010) on the consequences of 'relative age' (i.e., age differences within an age group, e.g., 15-year-olds) in sports and education.

extraordinary), and similarly for the unlucky players—all again stand to lose about 5 francs. Some might get lucky or unlucky twice in a row, but they're more likely to end up somewhere near the 5-franc mark, i.e., most of the lucky and unlucky players will regress to the mean.

- The same principle is at play when the observations come about in part through skill (or some other construct) and in part through chance. For instance, the most successful stock broker of the year 2032 is likely not to perform as well in the year 2033—even if the conditions on the stock market are comparable and the broker didn't start to rest on his laurels. The reason could simply be that he had more than his fair share of luck in 2032—you need some luck to come out on top—and wasn't as lucky in 2033. As a result, his performance in the next year is likely to be closer to the average performance (i.e., he's regressed to the mean of stock broker performance).

- If you administer a reading test to a group of learners one week and another reading test a couple of weeks later, you're likely to find that the very worst readers on the first test are still pretty poor readers on the second test (= the skill part), but their performance won't be as atrocious—it'll seem as though they've made some progress. Similarly, the best readers on the first test are likely to still be good readers on the second test, but their performance probably won't be as exceptional—it'll seem as though they've become worse.

  But this pattern can be explained in terms of measurement error: Even if none of the learners actually learnt or unlearnt something, you're likely to find such a pattern. The reason is that, if you obtained a dismal score, you're likely to be a pretty poor reader *and* to have had some bad luck—perhaps the topic of the reading test just wasn't suited for you, or you were coming down with the flu. A couple of weeks later, you might encounter a topic you know a thing or two about or you might be in better physical shape. Similarly, if you scored exceptionally well on the first test, you may have had some luck with the test's topic or with other circumstances, and these may not be as favourable next time round.

Exercise 9.1. A nationwide standardised maths test is administered to all 5th graders. It turns out that the classes with the highest mean test scores tend to be pretty small. One possible explanation is that small classes are more conducive to learning maths. Another explanation is that this finding is an artefact of measurement error. (These explanations aren't mutually exclusive.)

(a) Explain how measurement error can give rise to this finding.

(b) How could you tease the two explanations apart? ◇

Exercise 9.2 (Reading assignment)**.** Read Chapter 2 in Chambers (2017). There are no guiding questions for this text; it should be intelligible enough. But by way of preparing for it, try to answer these questions.

(a) You recruit 60 participants, aged 8–88. Half of them are assigned to the experimental group; the others to the control group (random assignment). You run a significance test comparing the mean age in both groups. What is, at most, the probability that you'll obtain a significant result (i.e., $p \leq 0.05$)?

(b) Each of your participants throws a fair six-sided dice. You run another significance test to check if there's a mean difference in the number of pips obtain in the control and in the intervention groups. (Evidently, the intervention doesn't make you throw dice any better.) What is, at most, the probability that you'll obtain a significant result (i.e., $p \leq 0.05$)?

(c) What do you know about the probability of observing *either* a significant age difference between the two groups, *or* a significant difference in the mean number of pips obtained, *or* two significant differences? ♢

10 Questionable research practices

10.1 A paradox

Sterling et al. (1995) inspected 563 articles in psychology journals (published in 1986–1987) in which significance tests were used to answer the research question. In 538 of them (96%), the researchers reported a significant result that confirmed their own hypothesis. In medical journals, the figure was lower but still pretty high (270/316, 85%).

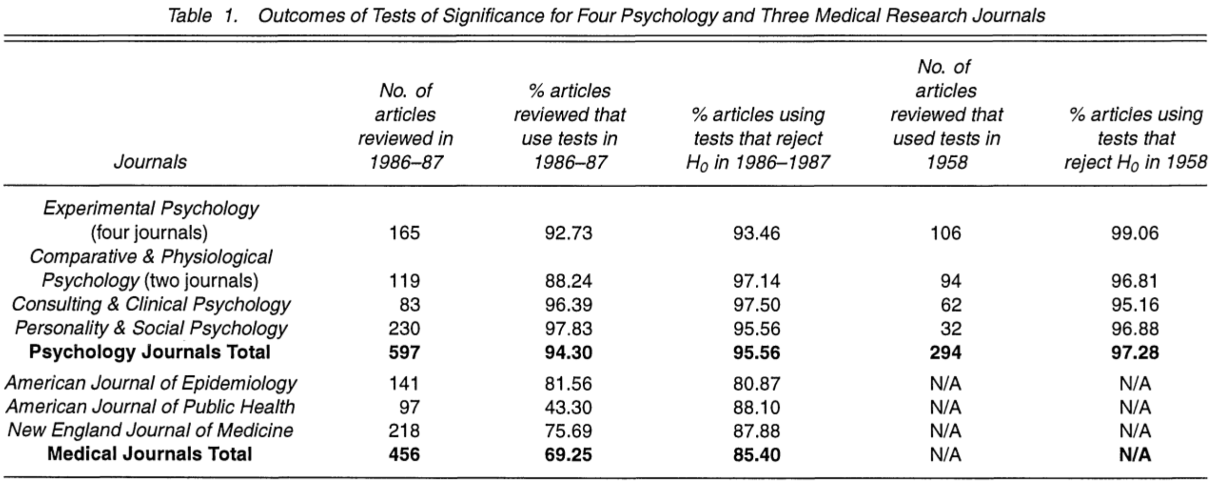


Table 1. Outcomes of Tests of Significance for Four Psychology and Three Medical Research Journals

Journals	*No. of articles reviewed in 1986–87*	*% articles reviewed that use tests in 1986–87*	*% articles using tests that reject H_0 in 1986–1987*	*No. of articles reviewed that used tests in 1958*	*% articles using tests that reject H_0 in 1958*
Experimental Psychology (four journals)	165	92.73	93.46	106	99.06
Comparative & Physiological Psychology (two journals)	119	88.24	97.14	94	96.81
Consulting & Clinical Psychology	83	96.39	97.50	62	95.16
Personality & Social Psychology	230	97.83	95.56	32	96.88
Psychology Journals Total	**597**	**94.30**	**95.56**	**294**	**97.28**
American Journal of Epidemiology	141	81.56	80.87	N/A	N/A
American Journal of Public Health	97	43.30	88.10	N/A	N/A
New England Journal of Medicine	218	75.69	87.88	N/A	N/A
Medical Journals Total	**456**	**69.25**	**85.40**	N/A	**N/A**

Figure 10.1: Table 1 from Sterling et al. (1995).

But at the same time, the sample sizes in psychological research are fairly small (Marszalek et al., 2011; Sedlmeier & Gigerenzer, 1989): The average study in applied psychology published in 1995 only contained 22 participants per condition. This implies that many of these studies must have had fairly low statistical power (see Chapter 4 on page 31): Even if the null hypothesis hadn't been correct in any of these studies, it'd have been impossible to reject it in 96% of cases.

For a long time (see already Sterling, 1959), it was believed that the reason for this discrepancy (low power, lots of significant results) was due to **publication bias**: Researchers prefer to write up the studies in which they obtained significant results, and editors and reviewers tend to reject studies with non-significant findings. The studies that were conducted but that produced non-significant findings were believed to languish in the researchers' file-drawers.

But while some studies never make it into print, the vast majority do. So where did the non-significant findings go?

10.2 *Hidden flexibility*

More recently, scholars with an interest in meta-science (i.e., science about science) have come to realise that research projects afford a great deal of flexibility. Researchers can—consciously or subconsciously—leverage this flexibility to produce a steady stream of significant findings—*even if the data are nothing but noise*.[1] Simmons et al. (2011) call this flexibility **researcher degrees of freedom** and superbly demonstrate how significant findings can be conjured from thin air if researchers afford themselves some leeway in analysing their data.

Sources of researcher degrees of freedom include:

- A researcher can run intermediate analyses and decide to stop or to continue collecing data based on the results. See Simmons et al. (2011) and Section 10.1 for the consequences of this.

- Sometimes, there are several ways in which a task or test can be scored, or how some variable can be constructed. When one way yields a significant finding and the other doesn't, it's easy to convince yourself that the one that produced significance was obviously the right one. Relatedly, researchers routinely collect multiple outcome variables, but it's tempting to focus on the one that 'worked' (i.e., produced significance) rather than on those that didn't. See Simmons et al. (2011), Gelman & Loken (2013), and von der Malsburg & Angele (2017). For a discussion with a focus on bilingualism research, see Poarch et al. (2019).

- **HARKing** (hypothesizing after the results are known; Kerr, 1998): A largely exploratory analysis is reported as though it were planned all along. Inevitably, the researchers will find in the data what they claim to have anticipated. (This can happen without any bad intent on the part of the researchers.)

- Convenient errors and biased debugging: Everyone makes mistakes, but you're more likely to catch your own mistakes when the results don't pan out than when they do. As a result, the mistakes that remain in the literature aren't distributed randomly but tend to favour the researchers' hypotheses.

Trying out several defensible analyses and glossing over the ones that didn't produce significance is referred to as **p-hacking**.

The practices listed above are examples of questionable research practices. Traditionally, these aren't viewed as outright fraud (which includes fabricating or manipulating data), though arguably,

[1] If the data aren't just noise, such flexibility will spuriously amplify the signal. For instance, even if A influences B, the literature as a whole will tend to overestimate the extent of this influence.

it will become increasingly difficult to invoke plausible deniability as professional researchers can be expected to know their consequences.

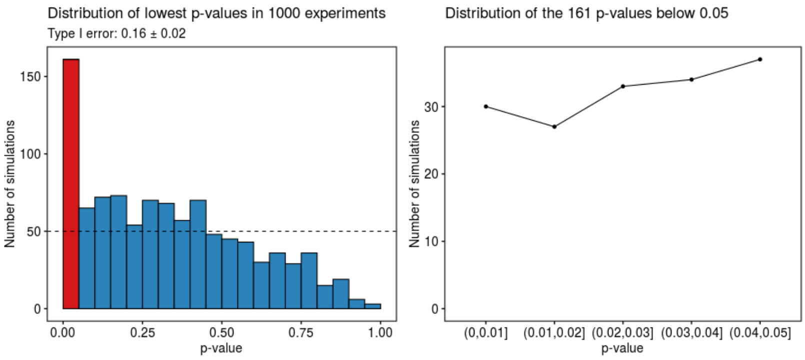
For possible solutions, see Chambers (2017).

Exercise 10.1 (False-positive psychology).

The consequences of researcher degrees of freedom/p-hacking are best appreciated by seeing them. Do these exercises in order.

(a) Open the app at `https://plurilinguisme.shinyapps.io/fppsy/` and carefully read the description.

(b) Click 'Simulate!', leaving all settings at their default values. Describe what the two graphs (reproduced here as Figure 10.2 for your convenience) are showing.


Distribution of lowest p-values in 1000 experiments
Type I error: 0.16 ± 0.02

Distribution of the 161 p-values below 0.05

Figure 10.2: When you run the app using its default settings, you'll obtain two graphs similar to these. Your graphs won't be identical as they are based on simulations with random data.

(c) First try to answer the following questions by *thinking* about them. Once you've written down your answer, check it by running the simulation.

i. Increase the 'maximum number of additional participants in each group' to 30. Leave the other settings at their default values. How will the graphs change?

ii. Leaving all other settings as they currently are, what will happen if instead of analysing the data after 10 new participants per condition, they're analysed after 5 new participants per condition? Or after just 2 new participants per condition?

iii. What'll happen when the correlation between the two dependent variables becomes weaker (e.g., $r = 0.1$ instead of $r = 0.5$)? Why?

iv. What'll happen when the correlation between the two dependent variables becomes stronger (e.g., $r = 0.95$)? Why?

v. For which combination of the different parameters will you obtain the highest Type-I error? Think before running the simulation!

vi. For which combination of the different parameters will you find a Type-I error rate of about 5%? Are there any parameters that don't play a role? Think before running the simulation! ◇

10.3 Optional: Further reading

Most studies referred to in this chapter are both accessible and short. If you read Simmons et al. (2011) (warmly recommended!), also read their short retrospective article (Simmons et al., 2018) lest you misinterpret the take-home message. Peterson (2016) presents an ethnographic study that gives you some insight into what questionable research practices look like in the field.

A highly accessible book-length treatment of these topics, and then some, which I cannot recommend highly enough, is Ritchie (2021). Chambers (2017) is also recommended.

A
Reading difficult results sections

Results sections in quantitative research reports can be daunting. Sometimes, the analyses are necessarily complex and require sophisticated knowledge about statistics and research design on the part of the reader. But too often, results sections are more difficult than they need to be (see Vanhove, 2021).

> Don't allow yourself to be dazzled by complicated analyses and incomprehensible results sections—the complexity may be largely superficial.

Here are some tips for muddling through difficult results sections with minimal psychological damage.[1]

1. Identify the central, genuine research questions and the corresponding hypotheses. Research papers in applied linguistics surprisingly often contain 'padding' research questions that are unrelated to the core goal of the study. When scanning the results section, you can usually leave aside the paragraphs about these uninteresting research questions. For example, in a report on a pretest/posttest experiment where participants were randomly assigned to conditions, you may find 'research' questions such as *Do participants in the treatment condition have different pretest scores from those in the control condition?* or *Do participants have higher scores on the posttest than on the pretest?* Both questions are uninteresting as they don't tell you whether the treatment actually worked.

2. **Draw a graph of the predictions.** (!) Having identified the key research questions and hypotheses, I often find it useful to sketch what the data would look like if the researchers' predictions panned out and what kind of data would, within reason, falsify their hypotheses. These graphs are usually simple hand-drawn line charts that illustrate the expected group differences. I find that they help me to better understand the logic behind the study and to focus on the important analyses in the Results section. You may find that several radically different patterns are in line with the authors' stated predictions; this tells you something

[1] Based on the blog entry *Surviving the ANOVA onslaught*.

about how specific their predictions are. (It's good to have specific as opposed to very general hypotheses!) It can also be useful to draw some simple graphs of data that would *not* be consistent with the authors' predictions. This, too, can help you work out if the predictions are fairly specific (a good thing) or if pretty much any pattern in the data would be consistent with them (a bad thing).

3. Look for a graph in the paper. Ideally, the paper will contain a graph of the main results that you can then compare with the graphs you drew yourself. Do the results seem to confirm or disconfirm the researchers' predictions? Sometimes, a good graph will allow you to carry out the easiest of significance tests yourself: the **inter-ocular trauma test**—if the conclusion hits you between the eyes, it's significant.[2] If the results are less clear cut, you'll need to scan the Results section for the more details, but by now, you should have a clearer idea of what you're looking for—and what you can ignore for now. If the paper doesn't contain a graph, you can often draw one yourself on the basis of the data provided in the tables.

4. Ignore tests unrelated to the central research questions. Results sections sometimes contain significance tests that are unrelated to the research questions the authors formulated up front (see Vanhove, 2021). Such tests include include balance tests in randomised experiments (e.g., "The control and intervention group did not differ in terms of SES ($t(36) = 0.8$, $p = 0.43$)."), tautological tests (e.g., "A one-way ANOVA confirmed that participants categorised as young, middle-aged and old differed in age ($F(2, 57) = 168.2$, $p < 0.001$).") as well as some less obvious cases. By and large, these tests tend to add little to the study. In non-randomised experiments, systematic differences on background variables between the groups may represent confounds, but these can be assessed based on the descriptive statistics and don't need to be rubber-stamped with a significance test.

Evidently, you'll get better at this with practice, and it'll be helpful to educate yourself on basic statistics, too. The latter will help you to understand better what was done, but also it will also allow you to ask more critical questions, not least of which is *Are these analyses at all relevant?*.

[2] To be clear, this isn't a formal significance test. But it's a useful heuristic!

B
Reporting research transparently

Many a research report leaves out information that is crucial for interpreting its findings correctly. And often, readers are implicitly asked to just take the authors word for it and trust that the analyses were run appropriately—even though reporting errors are common (Nuijten et al., 2016) and suboptimal or downright wrong analyses abound. Here are some tips to help ensure that your methods and findings are transparent to the readers.

1. You may be tempted to write long reports detailing everything. But such reports quickly become unreadable. My own preference is to aim for a crisp main text that doesn't inundate the reader with numbers and numbing details. Instead, I try to communicate the findings mainly through plots and refer to copious online materials for the details (Vanhove, 2021, also contains further guidance for writing quantitative research reports and some useful references). Using these online materials, interested readers should at least be able to the results occurring in the report, and so the online materials that I make available minimally comprise the data sets and the computer code necessary for reproducing the plots and numbers in the main text. Materials such as stimulus lists, questionnaires, code for running the experiment itself etc. should in my view also be shared by default. For projects involving lots of tedious steps that will be of little interest to the average reader, I also like to make available a technical report that documents every little detail (e.g., Vanhove et al., 2019).

   That said, **don't let perfect be the enemy of good**. If you're able to share your computer code but it's poorly documented, that's better than not sharing your code at all.

2. It's easier to share code, data, and materials if you made the decision to do so at the start of the project rather than a couple of weeks before handing in your report.

3. Nowadays I exclusively use `https://osf.io/` for making available online materials. See `https://osf.io/yxzfm/` for examples.

4. For further guidance, see Klein et al. (2018), and Levenstein & Lyle (2018, focus on sharing data), and Soderberg (2018, short tutorial on using osf.io).

References

Berthele, Raphael & Jan Vanhove. 2020. What would disprove interdependence? Lessons learned from a study on biliteracy in Portuguese heritage language speakers in Switzerland. *International Journal of Bilingual Education and Bilingualism* 23(5). 550–566. DOI: 10.1080/13670050.2017.1385590.

Bialystok, Ellen, Fergus I. M. Craik, Raymond Klein & Mythili Viswanathan. 2004. Bilingualism, aging, and cognitive control: Evidence from the Simon task. *Psychology and Aging* 19(2). 290–303. DOI: 10.1037/0882-7974.19.2.290.

Campbell, Donald T. & Julian C. Stanley. 1963. Experimental and quasi-experimental designs for research. In *Handbook of research on teaching*, Boston: Houghton Mifflin.

Chambers, Chris. 2017. *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton, NJ: Princeton University Press.

Christenfeld, Nicholas J. S., Richard P. Sloan, Douglas Carroll & Sander Greenland. 2004. Risk factors, confounding, and the illusion of statistical control. *Psychosomatic Medicine* 66. 868–875. DOI: 10.1097/01.psy.0000140008.70959.41.

Delsing, Lars-Olof & Katarina Lundin Åkesson. 2005. *Håller språket ihop Norden? En forskningsrapport om ungdomars förståelse av danska, svenska och norska*. Copenhagen: Nordiska ministerrådet.

Dewaele, Jean-Marc. 2008. The emotional weight of *I love you* in multilinguals' languages. *Journal of Pragmatics* 40(10). 1753–1780. DOI: 10.1016/j.pragma.2008.03.002.

Elwert, Felix. 2013. Graphical causal models. In Stephen L. Morgan (ed.), *Handbook of causal analysis for social research*, 245–273. Dordrecht, The Netherlands: Springer. DOI: 10.1007/978-94-007-6094-3_13.

Gelman, Andrew & Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time. URL `http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf`.

Goodman, Steven. 2008. A dirty dozen: Twelve *p*-value misconceptions. *Seminars in Hematology* 45. 135–140. DOI: 10.1053/j.seminhematol.2008.04.003.

Grosz, Michael P., Julia Rohrer & Felix Thoemmes. 2020. The taboo against explicit causal inference in nonexperimental psychology. *PsyArxiv* DOI: 10.31234/osf.io/8hr7n.

Helsen, Werner F., Jan van Winckel & A. Mark Williams. 2005. The relative age effect in youth soccer across Europe. *Journal of Sports Science* 23(6). 629–636. DOI: 10.1080/02640410400021310.

Huitema, Bradley E. 2011. *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies*. Hoboken, NJ: Wiley.

Imai, Kosuke, Gary King & Elizabeth A. Stuart. 2008. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171. 481–502. DOI: 10.1111/j.1467-985X.2007.00527.x.

Johnson, Daniel Ezra. 2013. Descriptive statistics. In Robert J. Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 288–315. Cambridge: Cambridge University Press.

Kahan, Brennen C., Sunita Rehal & Suzie Cro. 2015. Risk of selection bias in randomised trials. *Trials* 16. 405. URL http://www.trialsjournal.com/content/16/1/405.

Kang, Sean H. K., Tamar H. Gollan & Harold Pashler. 2013. Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin & Review* 20. 1259–1265. DOI: 10.3758/s13423-013-0450-z.

Keele, Luke, Randolph T. Stevenson & Felix Elwert. 2019. The causal interpretation of estimated associations in regression models. *Political Science Research and Methods* 8(1). 1–13. DOI: 10.1017/psrm.2019.31.

Kerr, Norbert L. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 2(3). 196–217. DOI: 10.1207/s15327957pspr0203_4.

Kirk, Neil W., Linda Fiala, Kennetch C. Scott-Brown & Vera Kempe. 2014. No evidence for reduced Simon cost in elderly bilinguals and bidialectals. *Journal of Cognitive Psychology* 26(4). 640–648. DOI: 10.1080/20445911.2014.929580.

Klein, Olivier, Tom E. Hardwicke, Frederik Aust, Johannes Breuer, Henrik Danielsson, Alicia Hofelich Mohr, Hans IJzerman, Gustav Nilsonne, Wolf Vanpaemel & Michael C. Frank. 2018. A practical guide for transparency in psychological science. *Collabra: Psychology* 4(1). 20. DOI: 10.1525/collabra.158.

Kolly, Marie-José. 2011. Weshalb hat man (noch) einen Akzent? Eine Untersuchung im Schnittfeld von Akzent und Einstellung bei Schweizer Dialektsprechern. *Linguistik Online* 50(6). 43–77.

Lardiere, Donna. 2006. Establishing ultimate attainment in a particular second language grammar. In ZhaoHong Han & Terrence Odlin (eds.), *Studies of fossilization in second language acquisition*, 35–55. Clevedon: Multilingual Matters.

Levenstein, Margaret C. & Jared A. Lyle. 2018. Data: Sharing is caring. *Advances in Methods and Practices in Psychological Science* 1(1). 95–103. DOI: 10.1177/2515245918758319.

Lorenzo, Francisco, Sonia Casal & Pat Moore. 2010. The effects of Content and Language Integrated Learning in European education: Key findings from the Andalusian Bilingual Sections Evaluation Project. *Applied Linguistics* 31(3). 418–442. DOI: 10.1093/applin/amp041.

Ludke, Karen M., Fernanda Ferreira & Katie Overy. 2014. Singing can facilitate foreign language learning. *Memory & Cognition* 42. 41–52. DOI: 10.3758/s13421-013-0342-5.

Marinova-Todd, Stefka H. 2011. "*Corplum* is a core from a plum": The advantage of bilingual children in the analysis of word meaning from verbal context. *Bilingualism: Language and Cognition* 15. 117–127. DOI: 10.1017/S136672891000043X.

Marszalek, Jacob M., Carolyn Barber & Julie Kohlhart. 2011. Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills* 112(2). 331–348. DOI: 10.2466/03.11.PMS.112.2.331-348.

McShane, Blakeley B, David Gal, Andrew Gelman, Christian Robert & Jennifer L Tackett. 2019. Abandon statistical significance. *The American Statistician* 73(Sup1). 235–245. URL `10.1080/00031305.2018.1527253`.

Mook, Douglas G. 1983. In defense of external invalidity. *American Psychologist* 38. 379–387. DOI: 10.1037/0003-066X.38.4.379.

Nuijten, Michèle B., Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp & Jelte M. Wicherts. 2016. The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods* 48(4). 1205–1226. DOI: 10.3758/s13428-015-0664-2.

Oehlert, Gary W. 2010. *A first course in the design and analysis of experiments*. URL `http://users.stat.umn.edu/~gary/book/fcdae.pdf`.

Oppenheimer, Daniel M., Tom Meyvis & Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45. 867–872. DOI: 10.1016/j.jesp.2009.03.009.

Peterson, David. 2016. The baby factory: Difficult research objects, disciplinary standards, and the production of statistical significance. *Socius* 2. 1–10. DOI: 10.1177/2378023115625071.

Poarch, Gregory J., Jan Vanhove & Raphael Berthele. 2019. The effect of bidialectalism on executive function. *International Journal of Bilingualism* 23(2). 612–628. DOI: 10.1177/1367006918763132.

Quené, Hugo. 2010. How to design and analyze language acquisition studies. In Elma Blom & Sharon Unsworth (eds.), *Experimental methods in language acquisition research*, 269–284. Amsterdam: John Benjamins.

Richardson, John T. E. 2018. The use of Latin-square designs in educational and psychological research. *Educational Research Review* 24. 84–97. DOI: 10.1016/j.edurev.2018.03.003.

Ritchie, Stuart. 2021. *Science fictions*. Penguin.

Rohrer, Julia M. 2018. Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science* 1(1). 27–42. DOI: 10.1177/2515245917745629.

Sedlmeier, Peter & Gerd Gigerenzer. 1989. Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* 105. 309–316.

Simmons, Joseph P., Leif D. Nelson & Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22. 1359–1366. DOI: 10.1177/0956797611417632.

Simmons, Joseph P., Leif D. Nelson & Uri Simonsohn. 2018. False-positive citations. *Perspectives on Psychological Science* 13(2). 255–259. DOI: 10.1177/1745691617698146.

Slavin, Robert E., Nancy Madden, Malgarita Calderón, Anne Chamberlain & Megan Hennessy. 2011. Reading and language outcomes of a multiyear randomized evaluation of transitional bilingual education. *Educational Evaluation and Policy Analysis* 33(1). 47–58. DOI: 10.3102/0162373711398127.

Slevc, L. Robert & Akira Miyake. 2006. Individual differences in second language proficiency: Does musical ability matter? *Psychological Science* 17(8). 675–681. DOI: 10.1111/j.1467-9280.2006.01765.x.

Soderberg, Courtney K. 2018. Using OSF to share data: A step-by-step guide. *Advances in Methods and Practices in Psychological Science* 1(1). 115–120. DOI: 10.1177/2515245918757689.

Sprietsma, Maresa. 2010. Effect of relative age in the first grade of primary school on long-term scholastic results: international comparative evidence using PISA 2003. *Education Economics* 18(1). 1–32. DOI: 10.1080/09645290802201961.

Sterling, Theodore D. 1959. Publication decision and their possible effect on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association* 54. 30–34.

Sterling, Theodore D., W. L. Rosenbaum & J. J. Weinkam. 1995. Publication decision revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician* 49. 108–112. URL http://www.jstor.org/stable/2684823.

Vanhove, Jan. 2015. Analyzing randomized controlled interventions: Three notes for applied linguists. *Studies in Second Language Learning and Teaching* 5. 135–152. DOI: 10.14746/ssllt.2015.5.1.7.

Vanhove, Jan. 2017. Lexical richness of short French, German and Portugese texts written by children (technical report). URL https://osf.io/vw4pc/.

Vanhove, Jan. 2020. Capitalising on covariates in cluster-randomised experiments. *PsyArXiv Preprints* DOI: 10.31234/osf.io/ef4zc.

Vanhove, Jan. 2021. Towards simpler and more transparent quantitative research reports. *ITL - International Journal of Applied Linguistics* 172(1). 3–25. DOI: 10.1075/itl.20010.van.

Vanhove, Jan & Raphael Berthele. 2017. Testing the interdependence of languages (HELASCOT project). In Raphael Berthele & Amelia Lambelet (eds.), *Heritage and school language literacy development in migrant children: Interdependence or independence?*, 97–118. Bristol: Multilingual Matters. DOI: 10.21832/9781783099054-007.

Vanhove, Jan, Audrey Bonvin, Amelia Lambelet & Raphael Berthele. 2019. Predicting perceptions of the lexical richness of short French, German, and Portuguese texts using text-based indices. *Journal of Writing Research* 10(3). 499–525. DOI: 10.17239/jowr-2019.10.03.04.

von der Malsburg, Titus & Bernhard Angele. 2017. False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language* 94. 119–133. DOI: 10.1016/j.jml.2016.10.003.

Westfall, Jacob & Tal Yarkoni. 2016. Statistically controlling for confounding constructs is harder than you think. *PLOS ONE* 11(3). e0152719. DOI: 10.1371/journal.pone.0152719.

Winter, Bodo. 2019. *Statistics for linguists: An introduction using R.* Routledge. DOI: 10.4324/9781315165547.