Reading 2: Designing the experiment and planning the sample

This reading is a selection from *Experimentology: An Open Science Approach to Experimental Psychology Methods* (Frank et al., 2025), prepared for LING 449U, University of Maryland, Fall 2026. The source, licence, and a statement of changes are at the end of the reading.

These excerpts cover conditions, repeated observations, statistical models, and sampling.

9.1.1 A two-factor experiment

Source

The classical "design of experiments" framework has as its goal to separate observed variability in the dependent measure into (1) variability due to the manipulation(s) and (2) other variability, including measurement error and participant-level variation. This framework maps nicely onto the statistical framework described in chapters 5–7. In essence, this framework models the distribution of the measure using the condition structure of our experiment as the predictor.

Different experimental designs will allow us to estimate specific effects more and less effectively. Recall in chapter 5, we estimated the effect of our tea/milk order manipulation by a simple subtraction: $\beta = \theta_T - \theta_C$ (where β is the effect estimate, and each θ indicates the estimates for a condition, treatment T and control C; we called them θ_T and θ_M in that chapter to denote tea- and milk-first conditions). This logic works just fine also if there are two distinct treatments in a three-condition experiment: each treatment can be compared to control separately. For treatment 1, $\beta_{T_1} = \theta_{T_1} - \theta_C$ and $\beta_{T_2} = \theta_{T_2} - \theta_C$.

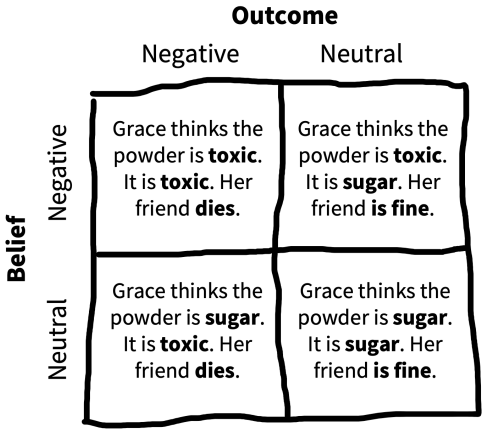


Figure 9.3: The 2 x 2 crossed design used in Young et al. (2007)

This logic is going to get more complicated if we have more than one distinct factor of interest, though. Let's look at an example.

Young et al. (2007) were interested in how moral judgments depend on both the beliefs of actors and the outcomes of their actions. They presented participants with vignettes in which they learned, for example, that Grace visits a chemical factory with her friend and goes to the coffee break room, where she sees a white powder that she puts in her friend's coffee. They then manipulated both Grace's *beliefs* and the *outcomes* of her action following the schema in figure 9.3. Participants ($N =$

10) used a four-point Likert scale to rate whether the actions were morally forbidden (1) or permissible (4). Figure 9.4 shows the data.

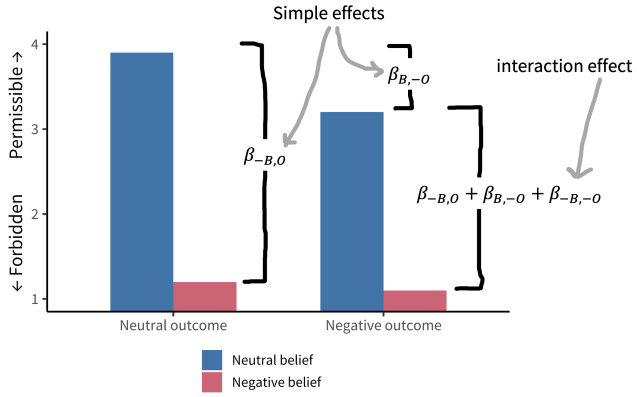


Figure 9.4: Moral permissibility as a function of belief and outcome. Results from Young et al. (2007), annotated with the estimated effects. Simple effects measure differences between the individual conditions and the neutral belief, neutral outcome condition. The interaction measures the difference between the predicted sum of the two simple effects and the actual observed data for the negative belief, negative outcome condition.

Young et al.'s design has two factors—belief and outcome—each with two levels (neutral and negative, noted as B and $-B$ for belief and O and $-O$ for outcome).[1] These factors are **fully crossed**: each level of each factor is combined with each level of each other.

This fully-crossed design makes it easy for us to estimate quantities of interest. Let's say that our **reference** group (equivalent to the control group for now) is neutral belief, neutral outcome. Now it's easy to use the same kind of subtraction we did before to estimate particular effects we care about. For example, we can look at the effect of negative belief in the case of a neutral outcome: $\beta_{-B,O} = \theta_{-B,O} - \theta_{B,O}$. This effect is shown on the left side of figure 9.4.

But now there is a complexity: these two **simple effects** (effects of one variable at a particular level of another variable) together suggest that the combined effect $\beta_{-B,-O}$ in the negative belief, negative outcome condition should be equal to the sum of $\beta_{-B,O}$ and $\beta_{B,-O}$.[2] As we can see from figure 9.4, that's not right. If it were, the negative belief, negative outcome condition would be below the lowest possible rating!

Instead, we observe an **interaction** effect (sometimes called a **two-way interaction** when there are two factors): the effect when both factors are present is different than the sum of the two simple effects. To capture this effect, we need an interaction term: $\beta_{-B,-O}$.[3] In other words, the effect of negative beliefs (intent) on subjective moral permissibility depends on whether the action caused harm. Critically, without a fully-crossed design, we can't estimate this interaction and we would have made an incorrect prediction about one condition.

[1] Neither of these is necessarily a "control" condition: the goal is simply to compare these two levels of the factor—negative and neutral—to estimate the effect due to the factor.

[2] If you're interested, you can also compute the **average** or **main** effect of a particular factor via the same subtractive logic. For example, the average effect of negative belief ($-B$) vs a neutral belief (B) can be computed as $\beta_{-B} = \frac{(\theta_{-O,-B} + \theta_{O,-B}) - (\theta_{-O,B} + \theta_{O,B})}{2}$.

[3] If you're reading carefully, you may be thinking that this sounds like we're talking about analysis of variance (ANOVA), not about experimental design per se. These two topics are actually the same topic! The question is how to design an experiment so that these statistical models can be used to estimate particular effects—and combinations of effects—that we care about. In case you missed it, we discuss modeling interactions in a regression framework in chapter 7.

9.1.2 Generalized factorial designs

Source

Young et al.'s design, in which there are two factors with two levels each, is called a **2 x 2 design** (pronounced "two by two"). These 2 x 2 designs are incredibly common and useful, but they are only one of an infinite variety of such designs that can be constructed.

Say we added a third factor to Young et al.'s design such that Grace either feels neutral toward her friend or is angry on that day. If we fully crossed this third factor A with the other two (B and O), we'd have a 2 x 2 x 2 design. This design would have eight conditions: (A, B, O), $(A, B, -O)$, $(A, -B, O)$, $(A, -B, -O)$, $(-A, B, O)$, $(-A, B, -O)$, $(-A, -B, O)$, $(-A, -B, -O)$. These conditions would allow us to estimate both two-way and three-way interactions (listed in table 9.1).

Table 9.1: Effects in a 2 x 2 x 2 design with affect, belief, and outcome as factors. Three-way interactions are hard to think about! The affect x belief x outcome interaction tells you about the difference in moral permissibility that's due to all three factors being present as opposed to what you'd predict on the basis of your estimates of the two-way interactions. In addition to being hard to think about, higher-order interactions tend to be hard to estimate, because estimating them accurately requires you to have a stable estimate of all of the lower-order interactions (McClelland and Judd 1993). For this reason, we recommend against experimental designs that rely on higher-order interactions unless you are in a situation where you both have strong predictions about these interactions and are confident in your ability to estimate them appropriately.

Things can get even more complicated. If you have three factors with two levels each, as in the example above (table 9.1), you can estimate seven total effects of interest. But if you have *four* factors with two levels each, you get 15. Four factors with *three* levels each gets you a horrifying 80 different effects![4] This way lies madness, at least from the perspective of estimating and interpreting individual effects in a reasonable sample size. Again, we suggest starting with one- and two-factor designs. There is a lot to be learned from simple designs that follow good measurement and sampling practices.

9.1.3 Between- vs within-participant designs

Source

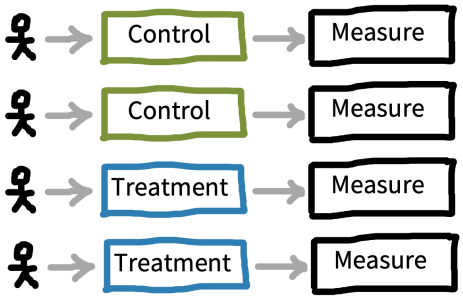


Figure 9.5: A between-participants design.

Once you know what factor(s) you would like to manipulate in your experiment, the next step is to consider how these will be presented to participants, and how that presentation will interact with your measurements. The biggest decision to be made is whether each participant will experience one level of a factor—a **between-participants** design—or whether they will experience multiple levels—a **within-participants** design. Figure 9.5 shows a simple example of between-participants design

[4]The general formula for N factors with M levels each is $M^N - 1$.

with four participants (two assigned to each condition), while figure 9.6 shows a within-participants version of the same design.

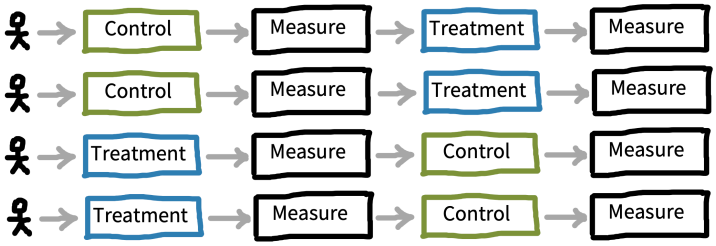


Figure 9.6: A within-participants design, counterbalanced for order (discussion of counterbalancing below).

Because people are very variable, the decision whether to measure a particular factor between- or within-participants is consequential. Imagine we're estimating our treatment effect as before, simply by computing $\hat{\beta} = \hat{\theta}_T - \hat{\theta}_C$ with each of these estimates from different populations of participants. In this scenario, our estimate $\hat{\beta}$ contains three components: (1) the true differences between θ_T and θ_C, (2) sampling-related variation in which participants from the population ended up in the samples for the two conditions, and (3) measurement error. Component 2 is present because any two samples of participants from a population will differ in their average on a measure—this is precisely the kind of sampling variation we saw in the null distributions in chapter 6.

When our experimental design is within participants, component 2 is not present because participants in both conditions are sampled from the *same* population. If we get unlucky and all of our participants are lower than the population mean on our measure, then that unluckiness affects our conditions equally. The consequences for choosing an appropriate sample size are fairly extreme: between-participants designs typically require between two and eight times as many participants as within-participants designs![5]

Given these advantages, why would you consider using a between-participants design? A within-participants design is simply not possible for all experiments. For example, consider a medical intervention—say, a new surgical procedure that is being compared to an established one. Patients cannot receive two different procedures, and so no within-participant comparison is possible.

Most manipulations in the behavioral sciences are not so extreme, but it still may be impractical or inadvisable to deliver multiple conditions. Greenwald (1976) distinguishes three types of undesirable effects:[6]

- **Practice effects** occur when administering the measure or the treatment will lead to change. Imagine a curriculum intervention for teaching a math concept—it would be hard to convince a school to teach the same topic to students twice, and the effect of the second round of teaching would likely be quite different than the first!
- **Sensitization effects** occur when seeing two versions of an intervention mean that you might respond differently to the second than the first because you have compared them and noticed the

[5] If you want to estimate how big an advantage you get from within-participants data collection, you need to know how correlated (reliable) your observations are. One analysis of this issue (Lakens 2016) suggests that the key relationship is $N_{\text{within}} = N_{\text{between}}(1 - \rho)/2$, where ρ is the correlation between the measurement of the two conditions within individuals. The more correlated they are, the smaller the within-participants N.

[6] We tend to think of all of these as being forms of carryover effect, and sometimes use this label as a catch-all description. Some people also use the picturesque description "poisoning the well" (Gelman 2017)—earlier conditions "ruin" the data for later conditions.

contrast. Consider a study on room lighting—if the experimenters are constantly changing the lighting, participants may become aware that lighting is the focus of the study!
- **Carryover effects** refer to the case where one treatment might have a longer-lasting effect than the measurement period. For example, imagine a study in which one treatment was to make participants frustrated with an impossible puzzle; if a second condition were given after this first one, participants might still be frustrated, leading to spillover of effects between conditions.

All of these issues can lead to real concerns with within-participant designs. But the desire for effect estimates that are completely unbiased by these concerns may lead to the overuse of between-participant designs (Gelman 2017). As mentioned above, between-participant designs come at a major cost in terms of power and precision.

An alternative approach is to acknowledge the possibility of carryover type effects and seek to mitigate them. First, you can make sure that the order of condition is randomized or balanced (see below); and second, you can analyze these carryover effects within your statistical model (for example by estimating the interaction of condition and order).[7]

We summarize the state of affairs from our perspective in figure 9.7. We think that within-participant designs should be preferred whenever possible. This conclusion is also consistent with meta-research we've done on replications from our course: across 176 student replications, the use of a within-subjects design was the strongest correlate of a successful replication (Boyce et al. 2023).[8]

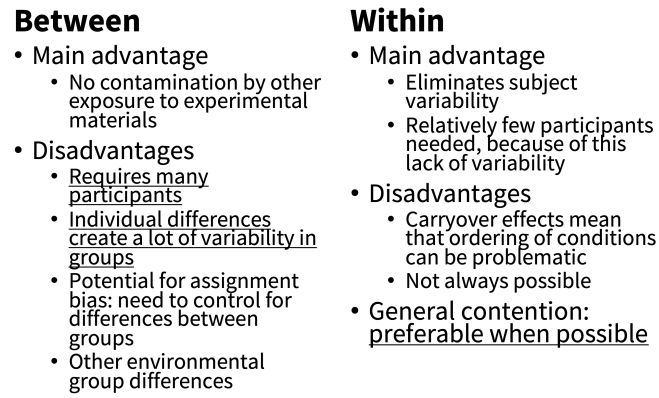


Figure 9.7: Pros and cons of between-participant vs within-participant designs. We recommend within-participant designs when possible.

9.1.4 Repeated measures and experimental items

Source

We just discussed decision-making about whether to administer multiple *manipulations* to a single participant. An exactly analogous decision comes up for *measures*! And our take-home will be similar: unless there are specific difficulties that come up, it's usually a very good idea to make multiple measurements (via multiple experimental **trials**) for each participant in each condition.

You can create a between-participants design where you administer your manipulation and then measure multiple times. This scenario is pictured in figure 9.8. Sometimes this works quite well. For example, imagine a transcranial magnetic stimulation (TMS) experiment: participants receive neural stimulation for a period of time, targeted at a particular region. Then they perform some measurement task repeatedly until it wears off. The more times they perform the measurement task,

[7]Even when one factor must be varied between participants, it is often still possible to vary others within subjects, leading to a mixed design in which some factors are between and others within.

[8]Caveat: the study design was observational, so no causal inference is possible.

the better the estimate of whatever effect (when compared to a control of TMS to another region, say).

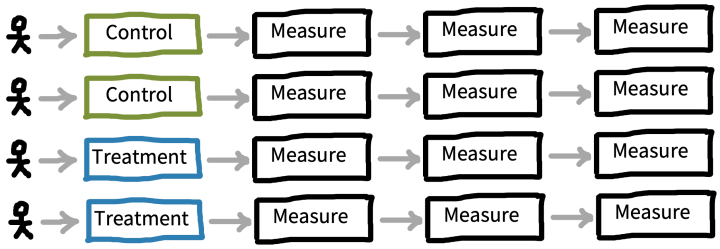


Figure 9.8: A between-participants, repeated-measures design.

Sometimes this design is called a **repeated measures** design, but terminology here is tricky again. The term "repeated measures" refers to any experiment where each participant is measured more than once, including both between-participants *and* within-participants designs.[9] Our advice is *both* to use within-participants designs *and* to get multiple measurements from each participant.

Why? In the last subsection, we described how variability in our estimates in a between-participants design depends on three components: (1) true condition differences; (2) sampling variation between conditions; and (3) measurement error.

Within-participants designs are good because they don't include (2). Repeated measures reduce (3): the more times you measure, the lower your measurement error, leading to greater measure reliability!

There are problems with repeating the same measure many times, however. Some measures can't be repeated without altering the response. To take an obvious example, we can't give the exact same math problem twice and get two useful measurements of mathematical ability! The typical solution to this problem is to create multiple items. In the case of a math assessment, you create multiple problems that you believe test the same concept but have different numbers or other superficial characteristics.

Using multiple items for measurement is good for two reasons. First, it reduces measurement error by allowing responses to be combined across items. But second, it increases the generalizability of the measurement. An effect that is consistent across many different items is more likely to be an effect that can be generalized to a whole class of stimuli—in precisely the same way that the use of multiple participants can license generalizations across a population of people (Clark 1973).

[9] We're talking about multiple trials with the same measure, not multiple distinct measures. As we discussed in chapter 8, we tend to be against measuring lots of different things in a single experiment—in part because of the concerns that we're articulating in this chapter: if you have time, it's better to make more precise measures of what you care about most. Measuring one thing well is hard enough. Much better to measure one thing well than many things badly.

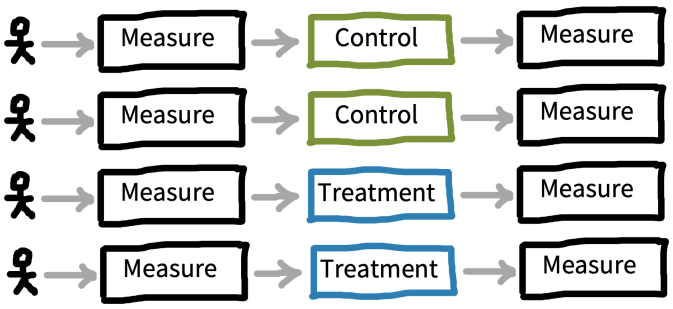


Figure 9.9: A between-participants, pre-post design.

One variation on the repeated measures, between-participants design is a specific version where the measure is administered both before (pre-) and after (post-) intervention, as in figure 9.9. This design is sometimes known as a **pre-post** design. It is extremely common in cases where the intervention is larger scale and harder to give within-participants, such as in a field experiment where a policy or curriculum is given to one sample and not to another. The premeasurements can be used to subtract out participant-level variability and recover a more precise estimate of the treatment effect. Recall that our treatment effect in a pure between-participants design is $\beta = \theta_T - \theta_C$. In a pre-post design, we can do better by computing $\beta = (\theta_{T_{\text{post}}} - \theta_{T_{\text{pre}}}) - (\theta_{C_{\text{post}}} - \theta_{C_{\text{pre}}})$. This equation says, "How much more did the treatment group go up than the control group?[10]

In sum, within-participants, repeated-measurement designs are the bread and butter of most research in perception, psychophysics, and cognitive psychology. When both manipulations and measures can be repeated, these designs afford high measurement precision even with small sample sizes; they are recommended whenever possible.

9.1.5 Discrete and continuous experimental manipulations

Source

Most experimental designs in psychology use discrete condition manipulations: treatment vs control. In our view, this decision often leads to a lost opportunity relative to a more continuous manipulation of the strength of the treatment. The goal of an experiment is to estimate a causal effect; ideally, this estimate can be generalized to other contexts and used as a basis for theory. Measuring not just one effect but instead a **dose-response** relationship—how the measure changes as the strength of the manipulation is changed—has a number of benefits in helping to achieve this goal.

Many manipulations can be **titrated**—that is, their strength can be varied continuously—with a little creativity on the part of an experimenter. A curriculum intervention can be applied at different levels of intensity, perhaps by changing the number of sessions in which it is taught. For a priming manipulation, the frequency or duration of prime stimuli can be varied. Two stimuli can be morphed continuously so that categorization boundaries can be examined.[11]

[10]This estimate is sometimes called a "difference in differences." The basic idea is widely used in the field of econometrics, both in experimental and quasi-experimental cases (Cunningham 2021). In practice, though, we recommend using the pre-treatment measurements as a covariate in a model-based analysis, not just doing the simple subtraction.

[11]These methods are extremely common in perception and psychophysics research, in part because the dimensions being studied are often continuous in nature. It would be basically impossible to estimate a participant's visual contrast sensitivity *without* continuously manipulating the contrast of the stimulus!

Dose-response designs are useful because they provide insight into the shape of the function mapping your manipulation to your measure. Knowing this shape can inform your theoretical understanding! Consider the examples given in figure 9.10.

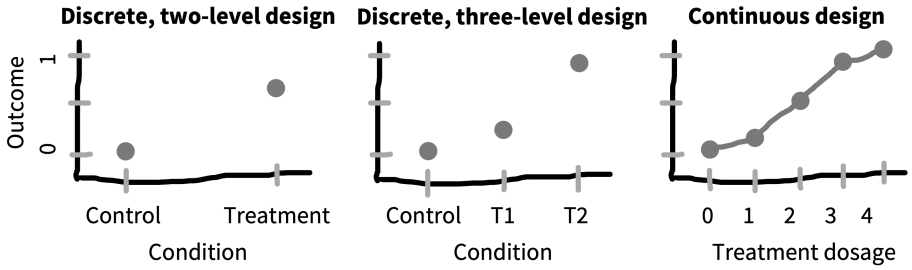


Figure 9.10: Three schematic designs. (left) Control and treatment are two levels of a nominal variable. (middle) Control is compared to ordered levels of a treatment. (right) Treatment level is an interval or ratio variable such that points can be connected and a parametric curve can be extrapolated.

If you only have two conditions in your experiment, then the most you can say about the relationship between your manipulation and your measure is that it produces an effect of a particular magnitude; in essence, you are assuming that condition is a nominal variable. If you have multiple ordered levels of treatment, you can start to speculate about the nature of the relationship between treatment and effect magnitude. But if you can measure the strength of your treatment, then you can begin to describe the nature of the relationship between the strength of treatment and strength of effect via a parametric function (e.g., a linear regression, a sigmoid, or other function).[12] These parametric functions can in turn allow you to generalize from your experiment, making predictions about what would happen under intervention conditions that you didn't measure directly!

9.2.1 Internal validity threats: Confounding

Source

First and foremost, manipulations must actually manipulate the construct whose causal effect is being estimated. If they *actually* manipulate something else instead, they are **confounded**. This term is used widely in psychology, but it's worth revisiting what it means. An **experimental confound** is a variable that is created in the course of the experimental design that is both causally related to the predictor and potentially also related to the outcome. As such, it is a threat to **internal validity**.

Let's go back to our discussion of causal inference in chapter 1. Our goal was to use a randomized experiment to estimate the causal effect of money on happiness. But just giving people money is a big intervention that involves contact with researchers—contact alone can lead to an experimental effect even if your manipulation fails. For that reason, many studies that provide money to participants either give a small amount of money or a large amount of money. This design keeps researcher contact consistent in both conditions, implying that the difference in outcomes between these two conditions should be due to the amount of money received (unless there are other confounds!).

Suppose you were designing an experiment of this sort and you wanted to follow our advice and use a within-participants design. You could measure happiness, give participants $100, wait a month and measure happiness again, give participants $1,000, wait a month, and then measure happiness for the third time. The trouble is, this design has an obvious experimental confound (figure 9.12): the order

[12]These assumptions are theory-laden, of course—the choice of a linear function or a sigmoid is not necessary: nothing guarantees that simple, smooth, or monotonic functions are the right ones. The key is that choosing a function makes explicit your assumptions about the nature of the treatment-effect relationship.

of the monetary gifts. Maybe happiness just went up more over time, irrespective of getting the second gift.

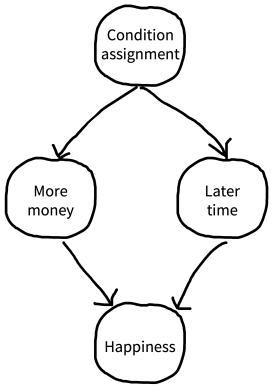


Figure 9.12: Confounding order and condition assignment means that you can't make an inference about the link between money and happiness.

If you think your experimental design might have a confound, you should think about ways to remove it. A first option is **elimination**, which we described above: basically, matching a particular variable across different conditions. This should be our first option for most confounds. Unfortunately, in our within-participants money-happiness study, order is confounded with condition so if we match orders we have eliminated our condition manipulation entirely.

A second option is **counterbalancing**, in which we vary a confounding factor systematically across participants so its average effect is zero across the whole experiment. In the case of our example, counterbalancing order across participants is a very safe choice. Some participants get $100 first and others get $1,000 first. That way, you are guaranteed that the order of conditions will have no effect of the confound on your average effect. The effect of this counterbalancing is that it "snips" the causal dependency between condition assignment and later time. We notate this on our causal diagram with a scissors icon (figure 9.13).[13] Time can still have an effect on happiness, but the effect is independent from the effect of condition and, hence, your experiment can still yield an unbiased estimate of the condition effect.

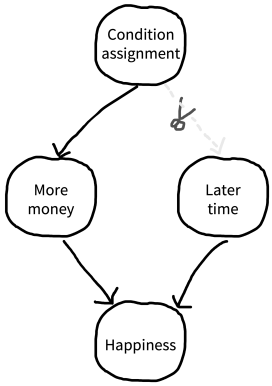


Figure 9.13: Confounding between a specific condition and the time of its administration can be removed by counterbalancing or randomization of order.

[13]In practice, counterbalancing is like adding an additional factor to your factorial design! But because the factor is a **nuisance factor**—basically, one we don't care about—we don't discuss it as a true condition manipulation. Despite that, it's a good practice to check for effects of these sorts of nuisance factors in your preliminary analysis. Even though your average effect won't be biased by it, it introduces variation that you might want to understand to interpret other effects and plan new studies.

Counterbalancing gets trickier when you have too many levels on a variable or multiple confounding variables. In that case, it may not be possible to do a full counterbalance so that all combinations of these factors are seen by equal numbers of participants. You may have to rely on partial counterbalancing schemes or Latin square designs (see the Depth box above; in this case, the Latin squares are used to create orderings of stimuli such that the position of each treatment in the order is controlled across two other confounding variables).

A final option, especially useful for such tricky cases, is **randomization**—that is, choosing which level of a nuisance variable to administer to the participant via a random choice. Randomization is increasingly common now that many experimental interventions are delivered by software. If you *can* randomize experimental confounds, you probably should. The only time you really get in trouble with randomization is when you have a large number of options, a small number of participants, or some combination of the two. Then you can end up with unbalanced levels of the randomized factors. Averaging across many experiments, a lack of balance will come out in the wash, but in a single experiment, it can lead to unfortunate bias in numbers.

A good approach to thinking through your experimental design is to walk through the experiment step by step and think about potential confounds. For each of these confounds, consider how it might be removed via counterbalancing or randomization. As our case study shows, confounds are not always obvious, especially in complex paradigms. There is no sure-fire way to ensure that you have spotted every one—sometimes the best way to avoid them is simply to present your candidate design to a skeptical friend.

9.2.2 Internal validity threats: Placebo, demand, and expectancy

Source

A second class of important threats to internal validity comes from cases where the research design is confounded by factors related to how the manipulation is administered, or even *that* a manipulation is administered. In some cases, these create confounds that can be controlled; in others they must simply be understood and guarded against. Rosnow and Rosenthal (1997) called these "artifacts": systematic errors related to research *on* people, conducted *by* people.

A **placebo** effect is a positive effect on the measure that comes as a result of participants' expectations about a treatment in the context of a research study. The classic example of a placebo is medical: giving an inactive sugar pill as a "treatment" leads some patients to report a reduction in whatever symptom they are being treated for. Placebo effects are a major concern in medical research as well as a fixture in experimental designs in medicine (Benedetti 2020). The key insight is that treatments must not simply be compared to a baseline of no treatment but rather to a baseline in which the psychological aspects of treatment are present but the "active ingredient" is not. In the terms we have been using, the experience of receiving a treatment (independent of the content of the treatment) is a confounding factor when you simply compare treatment to no treatment conditions.

If placebo effects reflect what participants expect from a treatment, then **demand characteristics** reflect what participants think *experimenters* want and their desire to help the experimenters achieve that goal (Orne 1962). Demand characteristics are often raised as a reason for avoiding within-participants designs—if participants become alert to the presence of an intervention, they may then respond in a way that they believe is helpful to the experimenter. Typical tools for controlling or identifying demand characteristics include using a cover story to mask the purpose of an experiment, using a debriefing procedure to probe whether participants typically guessed the purpose of an

experiment, and (perhaps most effectively) creating a control condition with similar demand characteristics but missing a key component of the experimental intervention. Note that if you use a cover story to mask the purpose of your experiment, it's worth thinking about whether you are using deception, which can raise ethical issues (see chapter 4). Certainly you should be sure to debrief participants about the true function of the experiment!

The final entry into this list of internal validity threats is **experimenter expectancy effects**, where the experimenter's behavior biases participants in a way that results in the appearance of condition differences where no true difference exists. The classic example of such effects is from the animal learning literature and the story of Clever Hans. Clever Hans was a horse who appeared able to do arithmetic by tapping out solutions with his hoof. On deeper investigation, it became apparent he was being cued by his trainer's posture (apparently without the trainer's knowledge) to stop tapping when the desired answer was reached. The horse knew nothing about math, but the experimenter's expectations were altering the horse's behavior across different conditions.

In any experiment delivered by human experimenters who know what condition they are delivering, condition differences can result from experimenters imparting their expectations. Table 9.2 shows the results of a meta-analysis estimating sizes of expectancy effects in a range of domains—the magnitudes are shocking. There's no question that experimenter expectancy is sufficient to "create" many interesting phenomena artifactually. The mechanisms of expectancy are an interesting research topic in their own right; in many cases expectancies appear to be communicated nonverbally in much the same way that Clever Hans learned (Rosnow and Rosenthal 1997).

Domain d r Example of type of study Laboratory interviews 0.14 0.07 Effects of sensory restriction on reports of hallucinatory experiences Reaction time 0.17 0.08 Latency of word associations to certain stimulus words Learning and ability 0.54 0.26 IQ test scores, verbal conditioning (learning) Person perception 0.55 0.27 Perception of other people's success Inkblot tests 0.84 0.39 Ratio of animal to human Rorschach responses Everyday situations 0.88 0.40 Symbol learning, athletic performance Psychophysical judgments 1.05 0.46 Ability to discriminate tones Animal learning 1.73 0.65 Learning in mazes and Skinner boxes Weighted mean 0.70 0.33 Unweighted mean 0.74 0.35 Median 0.70 0.33 Table 9.2: Magnitudes of expectancy effects in various domains. Based on Rosenthal (1994). In medical research, the gold standard is an experimental design where neither patients nor experimenters know which condition the patients are in.[14] Results from other designs are treated with suspicion because of their vulnerability to demand and expectancy effects. In psychology, the most common modern protection against experimenter expectancy is the delivery of interventions by a computer platform that can give instructions in a coherent and uniform way across conditions.

In the case of interventions that must be delivered by experimenters, ideally experimenters should be unaware of which condition they are delivering. On the other hand, the logistics of maintaining experimenter ignorance can be quite complicated in psychology. For this reason, many researchers opt for lesser degrees of control: for example, choosing to standardize delivery of an intervention via a script. These designs are sometimes necessary for practical reasons but should be scrutinized closely. "How can you rule out experimenter expectancy effects?" is an uncomfortable question that should be asked more frequently in seminars and paper reviews.

[14]These are commonly referred to as **double-blind** designs (though the term **masked** is now often preferred).

9.2.3 External validity of manipulations

Source

The goal of a specific experimental manipulation is to operationalize a particular causal relationship of interest. Just as the relationship between measure and construct can be more or less valid, so too can the relationship between manipulation and construct. How can you tell? Just like in the case of measures, there's no one royal road to validity. You need to make a validity argument (Kane 1992).[15]

For testing the effect of money on happiness, our manipulation was to give participants $1,000. This manipulation is clearly face valid. But how often do people just receive a windfall of cash, versus getting a raise at work or inheriting money from a relative? Is the effect caused by *having* the money, or *receiving* the money with no strings attached? We'd have to do more experiments to see what aspect of the money manipulation was most important. Even in straightforward cases like this one, we need to be careful about the breadth of the claims we make.

Sometimes validity arguments are made based on the success of the manipulation in producing some change in the measurement. In the implicit theory of mind case study we began with, the stimulus contained an animated Smurf character, and the argument was that participants took the Smurf's beliefs into account in making their judgments. This stimulus choice seems surprising—not only would participants have to track the implicit beliefs of other *people* but they would also have to be tracking the beliefs of depictions of nonhuman, animated characters. On the other hand, based on the success of the manipulation, the authors made an *a fortiori* argument: if people track even an animated Smurf's beliefs, then they *must* be tracking the beliefs of real humans.

Let's look at one last example to think more about manipulation validity. Walton and Cohen (2011) conducted a short intervention in which college students ($N = 92$) read about social belonging and the challenges of the transition to college and then reframed their own experiences using these ideas. This intervention led to long-lasting changes in grades and well-being. While the intervention undoubtedly had a basis in theory, part of our understanding of the validity of the intervention comes from its efficacy: sense of belonging *must* be a powerful factor if intervening on it causes such big changes in the outcome measures.[16] The only danger is when the argument becomes circular—a theory is correct because the intervention yielded a success, and the intervention is presumed to be valid because of the theory. The way out of this circle is through replication and generalization of the intervention. If the intervention repeatably produces the outcome, as has been shown in replications of the sense of belonging intervention (Walton et al. 2020), then the manipulation becomes an intriguing target for future theories. The next step in such a research program is to understand the limitations of such interventions (sometimes called **boundary conditions**).

The next selections connect the design to statistical models and the people and items sampled.

7.1.1 Regression for estimating a simple treatment effect

Source

Let's start with one of these special cases, namely estimating a treatment effect, β, in a two-group design. In chapter 5, we solved this exact challenge for the tea-tasting experiment. We applied a classical model in which the milk-first ratings are assumed to be normally distributed with mean

[15] One caveat is that the validity of a manipulation incorporates the validity of the manipulation *and* the measure. You can't have a good estimate of a causal effect if the measurement is invalid.

[16] On the other hand, if the manipulation *doesn't* produce a change in your measure, maybe the manipulation is invalid, but the construct still exists. Sense of belonging could still be important even if my particular intervention failed to alter it!

$\theta_{\text{M}} = \theta_{\text{T}} + \beta$ and standard deviation σ.[17]

Let's now write that model as a regression model—that is, as a model that predicts each participant's tea rating, Y_i, given that participant's treatment assignment, X_i. $X_i = 0$ represents the control (milk-first) group and $X_i = 1$ represents the treatment (tea-first) group.[18] Here, Y_i is the dependent variable and X_i is the independent variable. The subscripts i index the participants. To make this concrete, you can see some sample tea-tasting data (the first three observations from each condition) below (table 7.1), with the index i, the condition and its predictor X_i, and the rating Y.

Let's write this model more formally as a "linear regression of Y on X." Conventionally, regression models are written with β symbols for all parameters, so we'll now use $\beta_0 = \theta_M$ for the mean in the milk-first group and $\beta_1 = \theta_T - \theta_M$ as the average difference between the tea-first and milk-first groups. This β is a generalization of the one we were using to denote the causal effect above and in the previous two chapters.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Table 7.1: Example tea tasting data (first three observations from each condition). The term $\beta_0 + \beta_1 X_i$ is called the **linear predictor**, and it describes the expected value of an individual's tea rating, Y_i, given that participant's treatment group X_i (the single independent variable in this model). That is, for a participant in the control group ($X_i = 0$), the linear predictor is just equal to β_0, which is indeed the mean for the control group that we specified above. On the other hand, for a participant in the treatment group, the linear predictor is equal to $\beta_0 + \beta_1$, which is the mean for the treatment group that we specified. In regression jargon, β_0 and β_1 are **regression coefficients**, where β_1 represents the association of the independent variable X with the outcome Y.

The term ϵ_i is the **error term**, referring to random variation of participants' ratings around the group mean.[19] Note that this is a very specific kind of "error"; it does not include "error" due to bias, for example. Instead, you can think of the error terms as capturing the "error" that would be associated with predicting any given participant's rating based on just the linear predictor. If you predicted a control group participant's rating as β_0, that would be a good guess—but you still expect the participant's rating to deviate somewhat from β_0 (i.e., due to variability across participants beyond what is captured by their treatment groups). In our regression model, the linear predictor and error terms together say that participants' ratings scatter randomly (in fact, normally) around their group means with standard deviation σ. And that is exactly the same model we posited in chapter 5.[20]

[17] Here's a quick reminder that "model" here is a way of saying "set of assumptions about the data generating procedure." So saying that some equation is a "model" is the same as saying that we think this is where the data came from. We can "turn the crank"—that is, generate data through the process that's specified in those equations, such as by pulling numbers from a normal distribution with mean $\theta_{\text{T}} + \beta$ and standard deviation σ. In essence, we're committing to the idea that this process will give us data that are substantively similar to the ones we have already.

[18] Using 0 and 1 is known as **dummy coding** and allows us to interpret the parameter as the difference of the treatment group (tea-first) from the baseline (milk-first). There are many other ways to code categorical variables, with other interpretations. As a practical tip, be careful to check how your variables are coded before reporting anything!

[19] Formally, we'd write $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The tilde means "is distributed as," and what follows is a normal distribution with mean 0 and variance σ^2.

[20] You may be wondering why so much effort was put into building boutique solutions for these special cases when a unified framework was available the whole time. A partial answer is that the classical infrastructure of statistics was developed before computers were widespread, and these special cases were chosen because they were easy to work with "analytically" (meaning to work out all the math by hand, using values from big numerical tables). Now that we have computers with more flexible algorithms, the model-based perspective is more practical and accessible than it used to be.

Now we have the model. But how do we estimate the regression coefficients β_0 and β_1? The usual method is called **ordinary least squares (OLS)**. Here's the basic idea. For any given regression coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we would obtain different **predicted values**, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, for each participant. Some regression coefficient estimates will yield better predictions than others. Ordinary least squares estimation is designed to find the values of the regression coefficients that optimize these predictions, meaning that the predictions are as close as possible to participants' true outcomes, Y_i.

Figure 7.1 illustrates the tea-tasting data for each condition (the dots) along with the model predictions for each condition β_0 and $\beta_0 + \beta_1$ (blue lines). The gap between each data point and the corresponding predictions (the thing that OLS wants to minimize) is shown by the dotted lines.[21] These distances are sample estimates, called **residuals**, of the true errors (ϵ_i). The left-hand plot shows the OLS coefficient values—the ones that move the model's predictions as close as possible to the data points, in the sense of minimizing the total squared length of the dashed lines. The right-hand plot shows a substantially worse set of coefficient values.

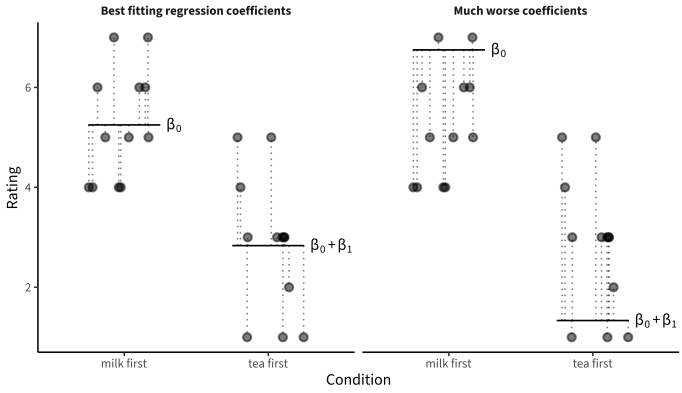


Figure 7.1: (left) Best-fitting regression coefficients for the tea-tasting experiment. (right) Much worse coefficients for the same data. Dotted lines: residuals. Circles: data points for individual participants.

You'll notice that we aren't talking much about p-values in this chapter. Regression models can be used to produce p-values for specific coefficients, representing inferences about the likelihood of the observed data under some null hypothesis regarding the coefficients. You can also compute Bayes Factors for specific regression coefficients, or use Bayesian inference to fit these coefficients under some prior expectation about their distribution. We won't talk much about this, or more generally how to fit the models we describe. Instead we want to help you begin thinking about making models of your data.

As it turns out, fitting an OLS regression model in R is extremely easy. The underlying function is `lm()`, which stands for linear model. You can fit the model with a single call to this function with a "formula" as its argument. Here's the call:

```
mod <- lm(rating ~ condition, data = tea_data)
```

[21] Ordinary least squares minimizes squared error loss, in the sense that it will choose the regression coefficient estimates whose predictions minimize $\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$, where n is the sample size. A wonderful thing about OLS is that those optimal regression coefficients (generically termed $\hat{\boldsymbol{\beta}}$) turn out to have a very simple closed-form solution: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. We are using more general notation here that supports multiple independent variables: $\hat{\boldsymbol{\beta}}$ is a vector, $\mathbf{X}$ is a matrix of independent variables for each subject, and $\mathbf{y}$ is a vector of participants' outcomes. As more good news, the standard error for $\hat{\boldsymbol{\beta}}$ has a similarly simple closed form!

Formulas in R are a special kind of terse notation for regression equations where you write the outcome, ~ (distributed as), and the predictors. R assumes that you want an intercept by default, and there are also a number of other handy defaults that make R formulas a nice easy way to specify relatively complex regression models, as we'll see below.

Once you've fit the model and assigned it to a variable, you can call `summary()` to see a summary of its parameters:

```
summary(mod)

Call:
lm(formula = rating ~ condition, data = tea_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8333 -1.2500  0.1667  0.7500  2.1667

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.2500     0.3688  14.234  1.4e-12 ***
conditiontea first  -2.4167     0.5216  -4.633 0.000129 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.278 on 22 degrees of freedom
Multiple R-squared:  0.4938,    Adjusted R-squared:  0.4708
F-statistic: 21.46 on 1 and 22 DF,  p-value: 0.0001286
```

You can also extract the coefficient values using `coef(mod)` and put them in a handy dataframe using `tidy(mod)` from the `broom` package (Robinson et al. 2023).

7.1.2 Adding predictors

Source

The regression model we just wrote down is the same model that underlies the t-test from chapter 6. But the beauty of regression modeling is that much more complex estimation problems can also be written as regression models that extend the model we made above. For example, we might want to add another predictor variable, such as the age of the participant.[22]

Let's add this new independent variable and a corresponding regression coefficient to our model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

Now that we have multiple independent variables, we've labeled them X_1 (treatment group) and X_2 (age) for clarity.

To illustrate how to interpret the regression coefficients in this model, let's use the linear predictor to compare the model's predicted tea ratings for two hypothetical participants who are both in the treatment group: 20-year-old Alice and 21-year-old Bob. Alice's linear predictor tells us that her

[22]The ability to estimate multiple coefficients at once is a huge strength of regression modeling, so much so that people sometimes use the label **multiple regression** to denote that there is more than one predictor + coefficient pair.

expected rating is $\beta_0 + \beta_1 + \beta_2 \cdot 20$. In contrast, Bob's linear predictor is $\beta_0 + \beta_1 + \beta_2 \cdot 21$. We could therefore calculate the expected difference in ratings for 21-year-olds versus 20-year-olds by subtracting Alice's linear predictor from Bob's, yielding just β_2.

We would get the same result if Alice and Bob were instead 50 and 51 years old, respectively. This equivalence illustrates a key point about linear regression models in general:

> The regression coefficient represents the expected difference in outcome when comparing any two participants who differ by 1 unit of the relevant independent variable, and who do not differ on any other independent variables in the model.

Here, the coefficient compares participants who differ by one year of age. In Section 7.4, we discuss whether and when to "control for" additional variables (i.e., when to add them to your model).

7.1.3 Interactions

Source

In our running example, we now have two predictors: condition and age. But what if the effect of condition varies depending on the age of the participant? This situation would correspond to a case where (say) older people were more sensitive to tea ordering, perhaps because of their greater tea experience. We call this an **interaction** effect: the effect of one predictor depends on the state of another.

Interaction effects are easily accommodated in our modeling framework. We simply add a term to our model that is the product of condition (X_1) and age (X_2), and weight this product by another beta, which represents the strength of this interaction:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$$

Statistical interactions are a very powerful modeling tool that can help us understand the relationship between different experimental manipulations or between manipulations and covariates (such as age). We discuss their role in experimental design—as well as some of the interpretive challenges that they pose—in much more detail in chapter 9.[23]

7.3 Linear mixed effects models

Source

Experimental data often contain multiple measurements for each participant (so-called **repeated measures**). In addition, these measurements are often based on a sample of stimulus items (which then each have multiple measures as well). This clustering is problematic for OLS models because the error terms for each datapoint are not independent.

Non-independence of datapoints may seem at first glance like a small issue, but it can present a deep problem for making inferences. Take the tea-tasting data we looked at above, where we had 24 observations in each condition. If we fit an OLS model, we observe a highly significant tea-first

[23] We won't go into this topic here, but we do want to provide a pointer to one of the most persistent challenges that comes up when you specify regression models with categorical predictors—and especially their interactions: how you "code" these categorical predictors. Above we created a "dummy" variable X that encoded milk-first tea as 0 and tea-first tea as 1. Dummy variables are very easy to think about, but in models with interactions, they can cause some problems. Because the interaction in our example model is a product of the dummy-coded condition variable and age, the interaction term β_3 is interpreted as the effect of age *for the tea-first condition* ($X = 1$) and hence the effect of age β_2 is actually the effect of age *for the milk-first condition.* The way to deal with this issue is to use a different coding system, such as **effects coding**. Davis (2010) gives a good tutorial on this tricky topic.

effect. Here is the estimate and confidence interval for that coefficient: $\hat{\beta} = -2.42$, 95% CI $[-3.50, -1.33]$. Based on what we talked about in the previous chapter, it seems like we'd be licensed in rejecting the null hypothesis that this effect is due to sampling variation and interpret this instead as evidence for a generalizable difference in tea preference in our sampled population.

But suppose we told you that all of those 48 total observations (24 in each condition) were from one individual named George. That would change the picture considerably. Now we'd have no idea whether the big effect we observed reflected a difference in the population, but we would have a very good sense of what George's preference is![24] The confidence intervals and p-values from our OLS model would be wrong now because all of the error terms would be highly correlated—they would all reflect George's preferences.

How can we make models that deal with clustered data? There are a number of widely used approaches for solving this problem including **linear mixed effects models**, **generalized estimating equations**, and **clustered standard errors** (often used in economics). Here we will illustrate how the problem gets solved in linear mixed models, which are an extension of OLS models that are fast becoming a standard in many areas of psychology (Bates et al. 2015).

7.3.1 Modeling random variation in clusters

In linear mixed effects models, we modify the linear predictor itself to model differences across clusters. Instead of just measuring George's preferences, suppose we modified the original tea-tasting experiment (without the age covariate) to collect ten ratings from each participant: five milk-first and five tea-first. We define the model the same way as we did before, with some minor differences:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_i + \epsilon_{it}$$

where Y_{it} is participant i's rating in trial t and X_{it} is the participant's assigned treatment in trial t (i.e., milk-first or tea-first).

If you compare this equation to the OLS equation above, you will notice that we added two things. First, we've added subscripts that distinguish trials from participants. But the big one is that we added γ_i, a separate intercept value for each participant. We call this a **random intercept** because it varies across participants (who are randomly selected from the population).[25]

The random intercept means that we have assumed that each participant has their own typical "baseline" tea rating—some participants overall just like tea more than others—and these baseline ratings are normally distributed across participants. Thus, ratings are correlated within participants because ratings cluster around each participant's *distinct* baseline tea rating. This model is better able to block misleading inferences. For example, suppose we only had one participant in each condition (say, George provided 24 milk-first ratings and Alice provided 24 tea-first ratings). If we found higher ratings in one condition, we would be able to attribute this difference to participant-level variation rather than to the treatment.[26]

[24] We discuss the strengths and weaknesses of repeated-measures designs like this in chapter 9 and the statistical trade-offs of having many people with a small number of observations per person vs a small number of people with many observations per person in chapter 10.

[25] Formally, we'd notate this random variation with $\gamma_i \sim N(0, \tau^2)$—in other words, that participants' random intercepts are sampled from a normal distribution around the shared intercept β_0 with standard deviation τ.

[26] Of course, this would be a terrible experiment! Ideally, we would address this problem upstream in our experiment design; see chapter 9.

Following the same logic, we could fit random intercepts for different stimulus items (for example, if we used different types of tea for different trials). We modeled participants as having normally distributed variation, and we can model stimulus variation the same way. Each stimulus item is assumed to produce a particular average outcome (i.e., some teas are tastier than others), with these average outcomes sampled from a normally distributed population.

Remarkably, linear mixed effects models are not much harder to specify in R than standard linear models. One very popular package is `lme4` (Bates et al. 2015), which provides the `lmer()` and `glmer()` functions (the latter for generalized linear mixed effects models). For our example here, we'd write:

To run the two mixed-model examples below, we first simulate a version of the tea data in which 12 participants each give three ratings in both conditions, so that each participant has an id.

```
tea_data <- expand_grid(id = 1:12, condition = c("milk first", "tea first"),
trial = 1:3) |>
  group_by(id) |>
  mutate(baseline = rnorm(1, sd = 0.75)) |>
  ungroup() |>
  mutate(rating = 3.5 + baseline + if_else(condition == "milk first", 1, 0) +
rnorm(n(), sd = 1),
         rating = pmin(pmax(round(rating), 1), 10)) |>
  select(id, condition, rating)

library(lme4)
lmer(rating ~ condition + (1 | id), data = tea_data)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: rating ~ condition + (1 | id)
   Data: tea_data
REML criterion at convergence: 212.3129
Random effects:
 Groups   Name        Std.Dev.
 id       (Intercept) 0.6760
 Residual             0.9372
Number of obs: 72, groups:  id, 12
Fixed Effects:
       (Intercept)  conditiontea first
            4.3333             -0.6389
```

In this model, the syntax `(1 | id)` specifies that we want a random intercept for each level of `id`.

7.3.2 Random slopes and the challenges of mixed effects models

Linear mixed effects models can be further extended to model clustering of the independent variables' *effects* within subjects, not just clustering of average *outcomes* within subjects. To do so, we can introduce **random slopes** (δ_i) to the model, which are multiplied by the condition variable X and represent differences across participants in the effect of tea tasting:

$$Y_i = \beta_0 + \beta_1 X_{it} + \gamma_i + \delta_i X_{it} + \epsilon_{it}$$

Just like the random intercepts, these random slopes will be assumed to vary across participants, following a normal distribution.[27]

This model now describes random variation in both overall how much someone likes tea *and* how strong their ordering preference is. Both of these likely do vary in the population, and so it seems like a good thing to put these in your model. Indeed, under some circumstances, adding random slopes is argued to be very important for making appropriate inferences.[28]

On the other hand, the model is much more complicated. When we had a simple OLS model above, we had only two parameters to fit (β_0 and β_1), but now we have those two plus two more, representing the standard deviations of the individual participant intercepts and slopes, plus parameters for each participant and for the condition effect for each participant. So we went from two parameters to 24![29] This complexity can lead to problems in fitting the models, especially with very small datasets (where these parameters are not very well constrained by the data) or very large datasets (where computing all these parameters can be tricky).[30]

Specifying random slopes in the `lme4` package is also relatively straightforward:

```
lmer(rating ~ condition + (condition | id), data = tea_data)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: rating ~ condition + (condition | id)
   Data: tea_data
REML criterion at convergence: 205.8406
Random effects:
 Groups   Name               Std.Dev. Corr
 id       (Intercept)        0.3787
          conditiontea first 0.6272   1.00
 Residual                    0.8744
Number of obs: 72, groups:  id, 12
Fixed Effects:
       (Intercept)  conditiontea first
            4.3333             -0.6389
optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 1 lme4
warnings
```

[27] These random slopes and intercepts can be assumed to be independent or correlated with one another, depending on the modeler's preference.

[28] There's lots of debate in the literature about the best random effect structure for mixed effects models. This is a very tricky and technical subject. In brief, some folks argue for so-called **maximal** models, in which you include every random effect that is justified by the design (Barr et al. 2013). Here that would mean including random slopes for each participant. The problem is that these models can get very complex and can be very hard to fit using standard software. We won't weigh in on this topic, but as you start to use these models on more complex experimental designs, it might be worth reading up on.

[29] Though we should note that these parameters aren't technically all independent from one another due to the structure of the mixed effects model.

[30] Many R users may be familiar with the widely used `lme4` package (Bates et al. 2015) for fitting mixed effects models using frequentist tools related to maximum likelihood. Such models can also be fit using Bayesian inference with the `brms` package (Bürkner 2021), which provides many powerful methods for specifying complex models.

Here, `(condition | id)` means "a separate random slope for `condition` should be fit for each level of `id`." Of course, specifying such a model is easier than fitting it correctly.

More generally, linear mixed effects models are very flexible, and they have become quite common in psychology. But they do have significant limitations. As we discussed, they can be tricky to fit in standard software packages. Further, the accuracy of these models relies on our ability to specify the structure of the random effects correctly.[31] If we specify an incorrect model, our inferences will be wrong! But it is sometimes difficult to know how to check whether your model is reasonable, especially with a small number of clusters or observations.

7.4.1 Modeling rules of thumb

Source

Our approach to statistical modeling is to start with a "default model" that is known in the literature as a **saturated model**. The saturated model of an experiment includes the full design of the experiment—all main effects and interactions—and nothing else. If you are manipulating a variable, include it in your model. If you are manipulating two, include them both and their interaction. If your design includes repeated measurements for participants, include a random effect of participant; if it includes experimental items for which repeated measurements are made, include random effect of stimulus.[32]

Don't include lots of other stuff in your default model. You are doing a randomized experiment, and the strength of randomized experiments is that you don't have to worry about confounding based on the population (see chapter 1). So don't put a lot of covariates in your default model—usually don't put in any![33]

This default saturated model then represents a simple summary of your experimental results. Its coefficients can be interpreted as estimates of the effects of interest, and it can be used as the basis for inferences about the relation of the experimental effect to the population using either frequentist or Bayesian tools.

Here's a bit more guidance about this modeling strategy.

1. **Preregister your model**. If you change your analysis approach after you see your data, you risk *p*-hacking—choosing an analysis that biases the estimate of your effect of interest. As we discussed in chapter 3 and as we will discuss in more detail in chapter 11, one important strategy for minimizing this problem is to **preregister** your analysis.[34]
2. **Visualize the model predictions against the observed data**. As we'll discuss in chapter 15, the "default model" for an experiment should go alongside a "default visualization" known as the **design plot** that similarly reflects the full design structure of the experiment and any primary

[31] One particularly problematic situation is when the correlation structure of the errors is mis-specified, for example if observations within a participant are more correlated for participants in the treatment group than in the control group; in such cases, mixed model estimates can be substantially biased (Bie et al. 2021).

[32] As discussed above, you can also include the maximal random effect structure (Barr et al. 2013), which involves random slopes as well as intercepts—but recognize that you cannot always fit such models.

[33] One corollary to having this kind of default perspective on data analysis: When you see an analysis that deviates substantially from the default, these deviations should provoke some questions. If someone drops a manipulation from their analysis, adds a covariate or two, or fails to control for some clustering in the data, did they deviate because of different norms in their subfield, or was there some other rationale? This line of reasoning sometimes leads to questions about the extent to which particular analytic decisions are post hoc and driven by the data (in other words, *p*-hacked). For an example, see the case study in chapter 11.

[34] A side benefit of preregistration is it makes you think through whether your experimental design is appropriate—that is, is there actually an analysis capable of estimating the effect you want from the data you intend to collect?

clusters. One way to check whether a model fits your data is then to plot it on top of those data. Sometimes this combination of model and data can be as simple as a scatter plot with a regression line. But seeing the model plotted alongside the data can often reveal a mismatch between the two. A model that does not describe the data very well is not a good source of generalizable inferences!

3. **Interpret the model predictions**. Once you have a model, don't just read off the p-values for your coefficients of interest. Walk through each coefficient, considering how it relates to your outcome variable. For a simple two-group design like we've been considering, the condition coefficient is the estimate of the causal effect that you intended to measure! Consider its sign, its magnitude, and its precision (standard error or confidence interval).

That said, there are some contexts in which it does make sense to depart from the default saturated model. For example, there may be insufficient statistical power to estimate multiple interaction terms, or covariates might be included in the model to help handle certain forms of missing data. The default model simply represents a very good starting point.

10.1 Sampling theory

Source

The basic idea of sampling is simple: you want to estimate some measurement for a large or infinite population by measuring a sample from that population.[35] Sampling strategies are split into two categories. **Probability sampling** strategies are those in which each member of the population has some known, prespecified probability of being selected to be in the sample—think "generalizing to Japanese people by picking randomly from a list of everyone in Japan." **Non-probability sampling** covers strategies in which probabilities are unknown or shifting, or in which some members of the population could never be included in the sample—think "generalizing to Germans by sending a survey to a German email list and asking people to forward the email to their family."

10.1.1 Classical probability sampling

In classical sampling theory, there is some **sampling frame** containing every member of the population—think of a giant list with every adult human's name in it. Then we use some kind of **sampling strategy**, maybe at the simplest just a completely random choice, to select N humans from that sample frame, and then we include them in our experiment. This scenario is the one that informs all of our statistical results about how sample means converge to the population mean (as in chapter 6).

Unfortunately, we very rarely do sampling of this sort in psychological research. Gathering true probability samples from the large populations that we'd like to generalize to is far too difficult and expensive. Consider the problems involved in doing some experiment with a sample of *all adult humans*, or even *adult English-speaking humans who are located in the United States*. As soon as you start to think about what it would take to collect a probability sample of this kind of population, the complexities proliferate. How will you find their names—what if they aren't in the registries you can access? How will you contact them—what if they don't have an email address? How will they do your experiment—what if they don't have an up-to-date web browser? What if they don't want to participate at all?

[35] There are some tools for dealing with estimation in smaller populations where your sample is a substantial fraction of the population (e.g., a survey of your department where you get responses from half of the students). We won't discuss those here; our focus is on generalizing to large populations of humans.

Instead, the vast majority of psychology research has been conducted with **convenience samples**: non-probability samples that feature individuals who can be recruited easily, such as college undergraduates or workers on crowdsourcing platforms like Amazon Mechanical Turk or Prolific Academic (see chapter 12). We'll turn to these below.

For survey research, on the other hand—think of election polling—there are many sophisticated techniques for dealing with sampling; although this field is still imperfect, it has advanced considerably in trying to predict complex and dynamic behaviors. One of the basic ideas is the construction of **representative samples**: samples that match the population in their representation of one or more sociodemographic characteristics like gender, income, race and ethnicity, age, or political orientation.

Representative samples can be constructed by probability sampling, but they can also be constructed through non-probability methods like recruiting quotas of individuals from different groups via various different convenience methods. These methods are critical for much social science research, but they have been used less frequently in experimental psychology research and aren't necessarily a critical part of the beginning experimentalist's toolkit.[36]

10.2.1 The worst version of the problem

Source

Psychology is the study of the human mind. But from a sampling theory standpoint, not a single estimate in the published literature is based on a simple random sample from the human population. And the situation is worse than that. Here are three of the most severe issues that have been raised regarding the generalizability of psychology research.

1. **Convenience samples**. Almost all research in experimental psychology is performed with convenience samples. This issue has led to the remark that "the existing science of human behavior is largely the science of the behavior of sophomores" (McNemar 1946, 333; quoted in Rosenthal and Rosnow 1984, 261). The samples we have easy access to just don't represent the populations we want to describe! At some point there was a social media account devoted to finding biology papers that made big claims about curing diseases and appending the qualifier "in mice" to them. We might consider whether we need to do the same to psychology papers. Would "Doing fonzy improves smoodling *in sophomore college undergraduates in the Western US*" make it into a top journal?

2. **The WEIRD problem**. Not only are the convenience samples that we study not representative of the local or national contexts in which they are recruited but those local and national contexts are also unrepresentative of the broad range of human experiences. Henrich et al. (2010) coined the term WEIRD (Western, educated, industrialized, rich, and democratic) to sum up some of the ways that typical participants in psychology experiments differ from other humans. The vast overrepresentation of WEIRD participants in the literature has led some researchers to suggest that published results simply reflect "WEIRD psychology"—a small and idiosyncratic part of a much broader universe of human psychology.[37]

[36]Readers can come up with counter-examples of recent studies that focus on representative sampling, but our guess is that they will prove the rule more generally. For example, a recent study tested the generality of growth mindset interventions for US high school students using a national sample (Yeager et al. 2019). This large-scale study sampled more than 100 high schools from a sampling frame of all registered high schools in the US, then randomly assigned students within schools that agreed to participate. They then checked that the schools that agreed to participate were representative of the broader population of schools. This study is great stuff, but we hope you agree that if you find yourself in this kind of situation—planning a multi-investigator five-year consortium study on a national sample—you might want to consult with a statistician and not use an introductory book like this one.

3. **The item sampling issue**. As we discussed in chapter 7 and 9, we're typically not just trying to generalize to new people; we're also trying to generalize to new stimuli (Westfall et al. 2015). The problem is that our experiments often use a very small set of items, constructed by experimenters in an ad hoc way rather than sampled as representatives of a broader population of stimuli that we hope to generalize to with our effect size estimate. What's more, our statistical analyses sometimes fail to take stimulus variation into account. Unless we know about the relationship of our items to the broader population of stimuli, our estimates may be based on unrepresentative samples in yet another way.

In sum, experiments in the psychology literature primarily measure effects from WEIRD convenience samples of people and unsystematic samples of experimental stimuli. Should we throw up our hands and resign ourselves to an ungeneralizable "science" of sample-specific anecdotes (Yarkoni 2020)?

10.2.2 Reasons for hope and ways forward

Source

We think the situation isn't as bleak as the arguments above might have suggested. Underlying each of the arguments above is the notion of **heterogeneity**, the idea that particular effects vary in the population.

Let's think through a very simple version of this argument. Say we have an experiment that measures the smoodling effect, and it turns out that smoodling is completely universal and invariant throughout the human population. Now, if we want to get a precise estimate of smoodling, we can take *any* sample we want because everyone will show the same pattern. Because smoodling is homogeneous, a nonrepresentative sample will not cause problems. There are some phenomena like this! For example, the Stroop task produces a consistent and similar interference effect for almost everyone (Hedge et al. 2018).

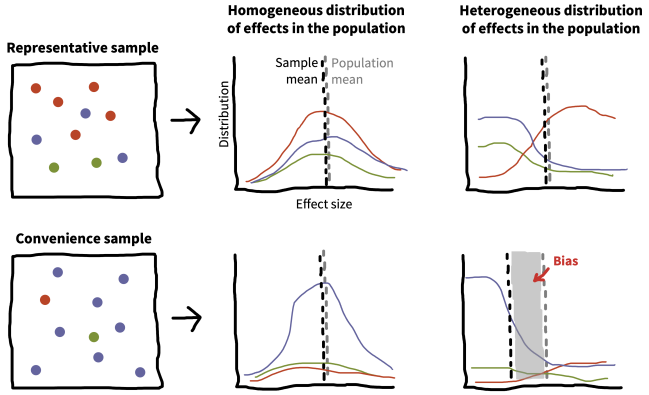


Figure 10.4: An illustration of the interaction of heterogeneity and convenience samples. Colors indicate arbitrary population subgroups. The left-hand panels show sample composition. Individual plots show the distribution of effect sizes in each subgroup.

Figure 10.4 illustrates this argument more broadly. If you have a representative sample (top), then your sample mean and your population mean will converge to the same value, regardless of whether the effect is homogeneous (right) or heterogeneous (right). That's the beauty of sampling theory. If you have a convenience sample, one part of the population is overrepresented in the sample. The convenience sample doesn't cause problems if the size of your effect is homogeneous in the

[37] The term WEIRD has been very useful in drawing attention to the lack of representation of the breadth of human experiences in experimental psychology. But one negative consequence of this idea has been the response that what we need to do as a field is to sample more "non-WEIRD" people. It is not helpful to suggest that every culture outside the WEIRD moniker is the same (Syed and Kathawalla 2020). A better starting point is to consider the way that cultural variation might guide our choices about sampling.

population—as with the case of smoodling or Stroop. The trouble comes when you have an effect that is heterogeneous. Because one group is overrepresented, you get systematic bias in the sample mean relative to the population mean.

So the problems listed above—convenience samples, WEIRD samples, and narrow stimulus samples—only cause issues if effects are heterogeneous. Are they? The short answer is, *we don't know*. Convenience samples are fine in the presence of homogeneous effects, but if we only use convenience samples, so we may not know which effects are homogeneous! Our metaphorical heads are in the sand.

We can't do better than this circularity without a theory of what should be variable and what should be consistent between individuals.[38] As naive observers of human behavior, differences between people often loom large. We are keen observers of social characteristics like age, gender, race, class, and education. For this reason, our intuitive theories of psychology often foreground these characteristics as the primary locus for variation between people. Certainly these characteristics are important, but they fail to explain many of the *in*variances of human psychology as well. An alternative line of theorizing starts with the idea that "lower-level" parts of psychology—like perception—should be less variable than "higher-level" faculties like social cognition. This kind of theory sounds like a useful place to start, but there are also counter-examples in the literature, including cases of cultural variation in perception (Henrich et al. 2010).

Multi-lab, multi-nation studies can help to address questions about heterogeneity, breaking the circularity we described above. For example, ManyLabs 2 systematically investigated the replicability of a set of phenomena across cultures (Klein et al. 2018), finding limited variation in effects between WEIRD sites and other sites. And in a study comparing a set of convenience and probability samples, Coppock et al. (2018) found limited demographic heterogeneity in another sample of experimental effects from across the social sciences.

So there are at least some cases where we don't have to worry as much about heterogeneity. More generally, such large-scale studies offer the possibility of measuring and characterizing demographic and cultural variation—as well as how variation itself varies between phenomena.

10.3.2 Attrition bias

Source

Attrition is when people drop out of your study. You should do everything you can to improve participants' experiences (see chapter 12), but sometimes—especially when a manipulation is onerous for participants or your experiment is longitudinal and requires tracking participants for some time—you will still have participants withdraw from the study.

Attrition on its own can be a threat to the generalizability of an experimental estimate. Imagine you do an experiment comparing a new very intense after-school math curriculum to a control curriculum in a sample of elementary school children over the course of a year. By the end of the year, suppose many of your participants have dropped out. The families who have stayed in the

[38] Many people have theorized about the ways that culture and language in general might moderate psychological processes (e.g., Markus and Kitayama 1991). What we're talking about is related but slightly different—a theory not of what's different but of when there should be any difference and when there shouldn't be. As an example, Tsai's (2007) "ideal affect" theory predicts that there should be more similarities in the distribution of actual affect across cultures, but that cultural differences should emerge in *ideal affect* (what people want to feel like) across cultures. This is a theory of when you should see homogeneity and when you should see heterogeneity.

study are likely those who care most about math. Even if you see an effect of the curriculum intervention, this effect may generalize only to children in families who love math.

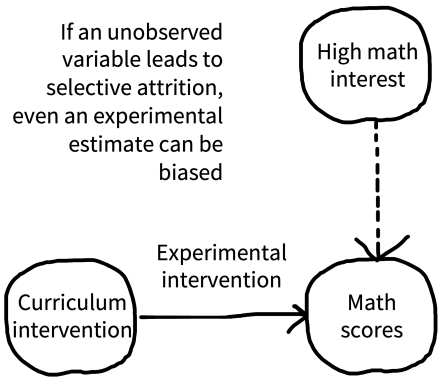


Figure 10.6: Selective attrition can lead to a bias even in the presence of random assignment. The dashed line indicates a causal relationship that is unobserved by the researcher.

But there is a further problem with attrition, known as **selective attrition**. If attrition is related to the outcome specifically within the treatment group (or for that matter, specifically within the control group), you can end up with a biased estimate, even in the presence of random assignment (Nunan et al. 2018). Imagine that students in the control condition of your math intervention experiment stayed in the sample, but the math intervention itself was so tough that most families dropped out except those who were very interested in math. Now, when you compare math scores at the end of the experiment, your estimate will be biased (figure 10.6): scores in the math condition could be higher simply because of differences in who stuck around to the end.[39]

Unfortunately, it turns out that attrition bias can be pretty common even in short studies, especially when they are conducted online when a participant can drop out simply by closing a browser window. This bias can be serious enough to lead to false conclusions. For example, Zhou and Fishbach (2016) ran an experiment in which they asked online participants to write about either four happy events (low difficulty) or 12 happy events (high difficulty) from the last year and then asked the participants to rate the difficulty of the task. Surprisingly, the high-difficulty task was rated as easier than the low-difficulty task!

Selective attrition was the culprit for this counterintuitive conclusion: while only 26% of participants dropped out of the low-difficulty condition, a full 69% dropped out of the high-difficulty task. The 31% that were left found it quite easy for them to generate 12 happy events, and so they rated the objectively harder task as less difficult.

Always try to track and report attrition information. That lets you—and others—understand whether attrition is leading to bias in your estimates or threats to the generalizability of your findings.[40]

Source and selection

Selected and rearranged passages from Michael C. Frank, Mika Braginsky, Julie Cachia, Nicholas A. Coles, Tom E. Hardwicke, Robert D. Hawkins, Maya B. Mathur, and Rondeline Williams (2025),

[39]If you get deeper into drawing DAGs like we are doing here, you will want to picture attrition as its own node in the graph, but that's beyond the scope of this book.

[40]If you get interested, there is a whole field of statistics that focuses on **missing data** and provides models for reasoning about and dealing with cases where data might not be **missing completely at random** (Little and Rubin 2019 is the classic reference for these tools). The causal inference frameworks referenced above also have very useful ways of thinking about this sort of bias.

Experimentology: An Open Science Approach to Experimental Psychology Methods. Stanford University. Book and citation. Also published by MIT Press.

The source book is licensed under Creative Commons Attribution-NonCommercial 4.0 International. The book's wording, section numbers, citations, and figure captions are retained. Sections have been rearranged; boxed digressions have been omitted; R code boxes are retained with the output of running them; margin notes appear as numbered footnotes; figures have been reduced. The annotated consent form is linked at full size. References to omitted material lead to the complete book. Source accessed November 5, 2026. This is an unofficial selection, not an author-approved edition.