Reading 1: From theory to a measurable prediction

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 move from causal questions and theory testing to measurement and uncertainty.

1.1.1 Describing causal relationships

Source

Consider the age-old question: Does money make people happy? This question is—at its heart—a question about what interventions on the world we can make. Can I get more money and make myself happier? Can I *cause* happiness with money?

How could we test our hypothesized effect of money on happiness? Intuitively, many people think of running an **observational study**. We might survey people about how much money they make and how happy they are. The result of this study would be a pair of measurements for each participant: `[money, happiness]`.

Now, imagine your observational study found that money and happiness were related—statistically **correlated** with each another: people with more money tended to be happier. Can we conclude that money causes happiness? Not necessarily. The presence of a correlation does not mean that there is a causal relationship!

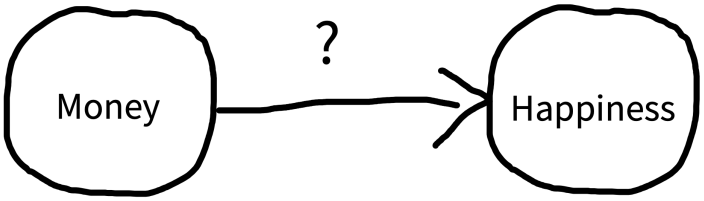


Figure 1.1: The hypothesized causal effect of money on happiness.

Let's get a bit more precise about our causal hypothesis. To illustrate causal relationships, we use a tool called **directed acyclic graphs** (DAGs, Pearl 1998). Figure 1.1 shows an example DAG for money and happiness. The arrow represents our idea about the potential causal link between two variables: money and happiness.[1] The direction of the arrow tells us which way we hypothesize that the causal relationship goes.

The correlation between money and happiness we saw in our observational study is consistent with the causal model in figure 1.1; however, it is also consistent with several alternative causal models, which we will illustrate with DAGs below.

[1] In this chapter, we're going to use the term "variables" without discussing why we study some variables and not others. In the next chapter, we'll introduce the term "construct," which indicates a psychological entity that we want to theorize about.

1.1.2 The problems of directionality and confounding

Source

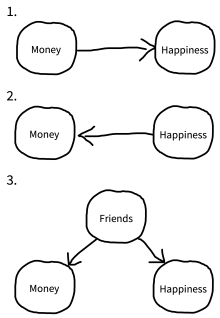


Figure 1.2: Three reasons why money and happiness can be correlated.

Figure 1.2 uses DAGs to illustrate several causal models that are consistent with the observed correlation between money and happiness. DAG 1 represents our hypothesized relationship—money causes people to be happy. But DAG 2 shows an effect in completely the opposite direction! In this DAG, being happy causes people to make more money.

Even more puzzling, there could be a correlation but no causal relationship between money and happiness in either direction. Instead, a third variable—often referred to as a **confound**—may be causing increases in both money and happiness. For example, maybe having more friends causes people to both be happier and make more money (DAG 3). In this scenario, happiness and money would be correlated even though one does not cause the other.

A confound (or several) may entirely explain the relationship between two variables (as in DAG 3), but it can also just *partly* explain the relationship. For example, it could be that money does increase happiness, but the causal effect is rather small, and only accounts for a small portion of the observed correlation between them, with the friendship confound (and perhaps others) accounting for the remainder.

In this case, because of the confounds, we say that the observed correlation between money and happiness is a **biased** estimate of the causal effect of money on happiness. The amount of bias introduced by the confounds can vary in different scenarios—it may only be small or it may be so strong that we conclude there's a causal relationship between two variables when there isn't one at all. The state of affairs summarized in figure 1.2 is why we say "correlation doesn't imply causation." A correlation between two variables *is consistent with* a causal relationship between them, but it's also consistent with other relationships.[2]

You can still learn about causal relationships from observational studies, but you have to take a more sophisticated approach. You can't just measure correlations and leap to causal conclusions. The "causal revolution" in the social sciences has been fueled by the development of statistical methods

[2]People sometimes ask whether *causation implies correlation* (the opposite direction). The short answer is "also no." A causal relationship between two variables often means that they will be correlated in the data, but not always. For example, imagine you measured the speed of a car and the pressure on the gas pedal / accelerator. In general, pressure and speed will be correlated, consistent with the causal relationship between the two. But now imagine you only measured these two variables when someone was driving the car up a hill—now the speed would be constant, but the pressure might be increasing, reflecting the driver's attempts to keep their speed up. So there would be no correlation between the two variables in that dataset, despite the continued causal relationship.

for reasoning about causal relationships from observational datasets.[3] As interesting as these methods are, however, they are only applicable in certain specific circumstances. In contrast, the experimental method *always* works to reduce bias due to confounding (though of course there are certain experiments that we can't do for ethical or practical reasons).

1.2 Experiments help us answer causal questions

Source

Imagine that you (a) created an exact replica of our world, (b) gave $1,000 to everybody in the replica world, and then (c) found a few years later that everyone in the replica world was happier than their matched self in the original world. This experiment would provide strong evidence that money makes people happier. Let's think through why.

Consider a particular person—if they are happier in the replica that in the original world, what could explain that difference? Since we have replicated the world exactly but made only one change—money—then that change is the only factor that could explain the difference in happiness. We can say that we *held all variables constant* except for money, which we **manipulated** experimentally, observing its effect on some **measure**—happiness. This idea—holding all variables constant except for the specific experimental manipulation—is the basic logic that underpins the experimental method (as articulated by Mill 1843).[4] Let's think back to our observational study of money and happiness. One big causal inference problem was the presence of "third-variable" confounds like having more friends. More friends could cause you to have more money and also cause you to be happier. The idea of an experiment is to hold everything else constant—including the number of friends that people have—so we can measure the effect of money on happiness. By holding the number of friends constant, we would be severing the causal links between friends and both money and happiness. This move is graphically conveyed in figure 1.3, where we "snip away" the friend confound.

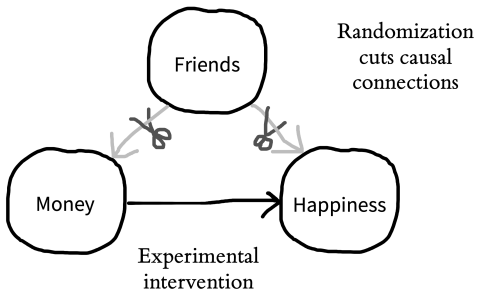


Figure 1.3: In principle, experiments allow us to "snip away" the friend confound by holding it constant (though in practice, it can be tough to figure out how to hold something constant when you are talking about people as your unit of study).

1.2.1 We can't hold people constant

This all sounds great in theory, you might be thinking, but we can't actually create replica worlds where everything is held constant, so how do we run experiments in the real world? If we were talking about experiments on baking cakes, it's easy to see how we could hold all of the ingredients

[3] In fact, DAGs are one of the key tools that social scientists use to reason about causal relationships. DAGs guide the creation of statistical models to estimate particular causal effects from observational data. We won't talk about these methods here, but if you're interested, check out the suggested readings at the end of this chapter.

[4] Another way to reason about why we can infer causality here follows the counterfactual logic we described in an earlier footnote. If the definition of causality is counterfactual ("What would have happened if the cause had been different?"), then this experiment fulfills that definition. In our impossible experiment, we can literally *see* the counterfactual: if the person had $1,000 more, here's how much happier they would be!

constant and just vary one thing, like baking temperature. Doing so would allow us to conduct an experimental test of the effect of baking temperature. But how can we "hold something constant" when we're talking about people? People aren't cakes. No two people are alike, and, as every parent with multiple children knows, even if you try to "hold the ingredients constant," they don't come out the same!

If we take two people and give one money, we're comparing two *different* people, not two instances of the same person with everything held constant. It wouldn't work to *make* the first person have more or fewer friends so they match the second person—that's not holding anything constant; instead it's another (big, difficult, and potentially unethical) intervention that might itself cause lots of effects on happiness.

You may be wondering: Why don't we just ask people how many friends they have and use this information to split them into equal groups? You could do that, but this kind of strategy only allows you to control for the confounds you know of. For example, you may split people equally based on their number of friends but not their educational attainment. If educational attainment also impacts both money and happiness, you still have a confound. You may then try to split people by both their number of friends and their education. But perhaps there's another confound you've missed: sleep quality! Similarly, it also doesn't work to select people who have the same number of friends—that only holds the friends variable constant and not everything *else* that's different between the two people. So what do we do instead?[5]

1.2.2 Randomization saves the day

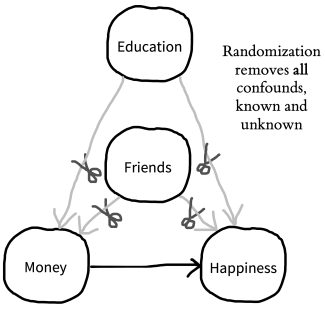


Figure 1.4: If you randomly split a large group of people into groups, the groups will, on average, be equal in every way.

The answer is **randomization**. If you randomly split a large roomful of people into two groups, the groups will, on average, have a similar number of friends. Similarly, if you randomly pick who in your experiment gets to receive money, you will find that the money and no-money groups, on average, have a similar number of friends. In other words, through randomization, the confounding role of friends is controlled. But the most important thing is that it's not *just* the role of friends that's controlled; educational attainment, sleep quality, and all the other confounds are controlled as well. If you randomly split a large group of people into groups, the groups will, on average, be equal in every way (figure 1.4).

So, here's our simple experimental design: we randomly assign some people to a money group and some people to a no-money control group (we sometimes call these groups **conditions**). Then we

[5]Many researchers who have seen regression models used in the social sciences assume that "controlling for lots of stuff" is a good way to improve causal inference. Not so! In fact, inappropriately controlling for a variable in the absence of a clear causal justification can actually make your effect estimate *more* biased (Wysocki et al. 2022).

measure the happiness of people in both groups. The basic logic of randomization is that, if money causes happiness, we should see more happiness—on average—in the money group.[6]

Randomization is a powerful tool, but there is a caveat: it doesn't work every time. *On average*, randomization will ensure that your money and no-money groups will be equal with respect to confounds like number of friends, educational attainment, and sleep quality. But just as you can flip a coin and sometimes get heads nine out of ten times, sometimes you use randomization and still get more highly educated people in one condition than the other. When you randomize, you guarantee that, on average, all confounds are controlled. Hence, there is no systematic bias in your estimate from these confounds. But there will still be some noise from random variation.

In sum, randomization is a remarkably simple and effective way of holding everything constant besides a manipulated variable. In doing so, randomization allows experimental psychologists to make unbiased estimates of causal relationships. Importantly, randomization works both when you do have control of every aspect of the experiment—like when you are baking a cake—and even when you don't—like when you are doing experiments with people.[7]

2.1.1 Psychological constructs

Source

Constructs are the psychological variables that we want our theory to describe, like "money" and "happiness" in the example from last chapter. At first glance, it might seem odd that we need a specific name for these variables. But in probing the relationship between money and happiness, we will have to figure out a way to measure happiness. Let's say we just ask people to answer the question "How happy are you?" by giving ratings on a 1 (miserable) to 10 (elated) scale.

Now say someone in the study reports they are an 8 on this scale. Is this *really* how happy they are? What if they weren't concentrating very hard on the rating, or if they thought the researcher wanted them to be happy? What if they act much less happy in their interactions with family and friends?

We resolve this dilemma by saying that the self-report ratings we collect are only a **measure** of a **latent construct**, happiness. The construct is latent because we can never see it directly, but we think it has a causal influence on the measure: happier people should, on average, provide higher ratings. But many other factors can lead to noise or bias in the measurement, so we shouldn't mistake those ratings as actually *being* the construct.

The particular question "How happy are you?" is one way of going from the general construct to a specific measure. The general process of going from construct to a specific instantiation that can be measured or manipulated is called **operationalization**. Happiness can be operationalized by self-report, but it can also be operationalized many other ways—for example, through a measure like the use of positive language in a personal essay, or by ratings by friends, family, or a clinician. These decisions about how to operationalize a construct with a particular measure are tricky and consequential, and we discuss them extensively in chapter 8. Each different operationalization might

[6] You may already be protesting that this experiment could be done better. Maybe we could measure happiness before and after randomization, to increase precision. Maybe we need to give a small amount of money to participants in the control condition to make sure that participants in both conditions interact with an experimenter, and hence, that the conditions are as similar as possible. We agree! These are important parts of experimental design, and we'll touch on them in subsequent chapters.

[7] There's an important caveat to this discussion: you don't always have to randomize *people*. You can use an experimental design called a **within-participants** design, in which the same people are in multiple conditions. This type of design has a different set of unknown confounds, this time centering around *time*. So, to get around them, you have to randomize the order in which your manipulation is delivered. This randomization works very well for some kinds of manipulations but not so well for others. We'll talk more about these kinds of designs in chapter 9.

be appropriate for a specific study, yet it would require some justification and argument to connect each one to the others.

Proposing a particular construct is a very important part of making a theory. For example, a researcher might worry that self-reported happiness is very different than someone's well-being as observed by the people around them and assert that happiness is not a single construct but rather a group of distinct constructs. This researcher would then be surprised to know that self-reports of happiness relate very highly to others' perceptions of a person's well-being (Sandvik et al. 1993).[8]

Even external, apparently nonpsychological variables like money don't have direct effects on people but rather operate through psychological constructs. People studying money seriously as a part of psychological theories think about perceptions of money in different ways depending on the context. For example, researchers have written about the importance of how much money you have on hand based on when in the month your paycheck arrives (Ellwood-Lowe et al. 2022) but have also considered perceptions of long-term accumulation of wealth as a way of conceptualizing people's understanding of the different resources available to White and Black families in the United States (Kraus et al. 2019).

Finally, a construct can be operationalized through a manipulation: in our money-happiness example, we operationalized "more money" in our theory with a gift of a specific amount of cash. We hope you see through these examples that operationalization is a huge part of the craft of being a psychology researcher—taking a set of abstract constructs that you're interested in and turning them into a specific experiment with a manipulation and a measure that tests your causal theory. We'll have a lot more to say about how this is done in chapter 9.

2.1.2 The relationships between constructs

Source

Constructs gain their meaning in part via their own definitions and operationalizations, but also in part through their causal relationships to other constructs. Figure 2.1 shows a schematic of what this kind of theory might look like—as you can see, it looks a lot like the DAGs that we introduced in the last chapter! That's no accident. The arrows here also describe hypothesized causal links.[9]

[8]Sometimes positing the construct *is* the key part of a theory. *g* (general intelligence) is the classic psychological example of a single-construct theory. The idea behind *g* theory is that the best measure of general intelligence is the shared variance between a wide variety of different tests. The decision to theorize about and measure a single unified construct for intelligence—rather than say, many different separate kinds of intelligence—is itself a controversial move.

[9]Sometimes these kind of diagrams are used in the context of a statistical method called structural equation modeling, where circles represent constructs and lines represent their relationships with one another. Confusingly, structural equation models are also used by many researchers to describe psychological theories. The important point for now is that they are one particular statistical formalism, not a general tool for theory building—the points we are trying to make here are more general.

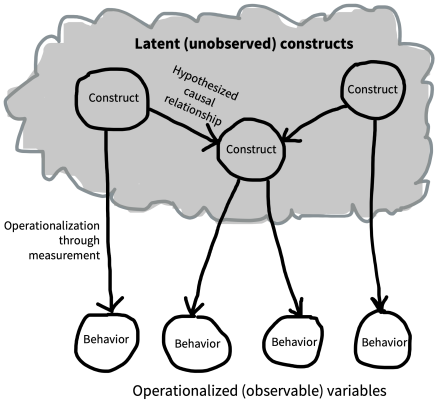


Figure 2.1: A schematic of what a theory might look like.

This web of constructs and assumptions is what Cronbach and Meehl (1955) referred to as a "nomological network"—a set of proposals about how different entities are connected to one another. The tricky part is that the key constructs are never observed directly. They are in people's heads.[10] So researchers only get to probe them by measuring them through specific operationalizations.

One poetic way of thinking about this idea is that the theoretical system of constructs "floats, as it were, above the plane of observation and is anchored to it by rules of interpretation" (Hempel 1952, 36). So, even if your theory posits that two constructs (say, money and happiness) are directly related, the best you can do is manipulate one operationalization and measure another operationalization. If this manipulation doesn't produce any effect, it's possible that you are wrong and money does not cause happiness—but it is also possible that your operationalizations are poor.

Here's a slightly different way of thinking about a theory. A theory provides a **compression** of potentially complex data into much a smaller set of general factors. If you have a long sequence of numbers, say [2 4 8 16 32 64 128 256 ...], then the expression 2^n serves as a compression of this sequence—it's a short expression that tells you what numbers are in vs out of the sequence. In the same way, a theory can compress a large set of observations (maybe data from many experiments) into a small set of relationships between constructs. Now, if your data are noisy, say [2.2 3.9 8.1 16.1 31.7 ...], then the theory will not be a perfect representation of the data. But it will still be useful.

In particular, having a theory allows you to **explain** observed data and **predict** new data. Both of these are good things for a theory to do. For example, if it turned out that the money causes happiness theory was true, we could use it to explain observations such as greater levels of happiness among wealthy people. We could also make predictions about the effects of policies like giving out a universal basic income on overall happiness.[11] Explanation is an important feature of good theories, but it's also easy to trick yourself by using a vague theory to explain a finding post hoc (after the fact). Thus, the best test of a theory is typically a new prediction, as we discuss below.

[10] We're not saying these should correspond to specific brain structures. They could, but most likely they won't. The idea that psychological constructs are not the same as any particular brain state (and especially not any particular brain region) is called "multiple realizability" by philosophers, who mostly agree that psychological states can't be reduced to brain states, as much as philosophers agree on anything (Block and Fodor 1972).

[11] The relationship between money and happiness is actually much more complicated than what we're assuming here. For example, Killingsworth et al. (2023) describes a collaboration between two sets of researchers who had different viewpoints on the connection between money and happiness.

One final note: Causal diagrams are a very useful formalism, but they leave the generalizability of the causal relationships implicit. For example, will more money result in more happiness for everyone, or just for people at particular ages or in particular cultural contexts? "Who does this theory apply to?" is an important question to ask about any proposed causal framework.

2.1.3 Specific theories vs general frameworks

Source

You may be thinking, "Psychology is full of theories but they don't look that much like the ones you're talking about!" Very few of the theories that bear this label in psychology describe causal relationships linking clearly defined and operationalized constructs. You also don't see that many DAGs, though these are getting (slightly) more common lately (Rohrer 2018).

Here's an example of something that gets called a theory yet doesn't share the components described above. Bronfenbrenner's (1992) ecological systems theory (EST) is pictured in figure 2.2. The key thesis of this theory is that children's development occurs in a set of nested contexts that each affect one another and in turn affect the child. This theory has been immensely influential. Yet, if it's read as a causal theory, it's almost meaningless: nearly everything connects to everything in both directions, and the constructs are not operationalized—it's very hard to figure out what kind of predictions it makes!

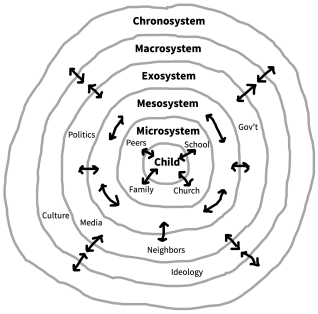


Figure 2.2: The diagram often used to represent Bronfenbrenner's ecological systems theory. Note that circles no longer denote discrete constructs; arrows can be interpreted as causal relationships, but all constructs are assumed to be fully connected.

Ecological systems theory is not really a theory in the sense that we are advocating for in this chapter—and the same goes for many other very interesting ideas in psychology. It's not a set of causal relationships between constructs that allow specific predictions about future observations. Ecological systems theory is instead a broad set of ideas about what sorts of theories are more likely to explain specific phenomena. For example, it helps remind us that a child's behavior is likely to be influenced by a huge range of factors, such that any individual theory cannot just focus on an individual factor and hope to provide a full explanation. In this sense, EST is a framework: it guides and inspires specific theories—in the sense we've discussed here, namely a set of causal relationships between constructs—without being a theory itself.

Frameworks like EST are often incredibly important. They can also make a big difference to practice. For example, EST supports a model in social work in which children's needs are considered not only as the expression of specific internal developmental issues but also as stemming from a set of overlapping contextual factors (Ungar 2002). Concretely, a therapist might be more likely to examine family, peer, and school environments when analyzing a child's situation through an EST lens.

There's a continuum between precisely specified theories and broad frameworks. Some theories propose interconnected constructs but don't specify the relationships between them, or don't specify

how those constructs should be operationalized. So when you read a paper that says it proposes a "theory," it's a good idea to to ask whether it describes specific relations between operationalized constructs. If it doesn't, it may be more of a framework than a theory.

2.2.2 A holistic viewpoint on theory testing

Source

The key issue that leads us to reject strict falsificationism is the observation that no individual hypothesis (a part of a theory) can be falsified independently. Instead, a large series of what are called **auxiliary assumptions** (or auxiliary hypotheses) are usually necessary to link an observation to a theory (Lakatos 1976). For example, if giving some individual person money didn't change their happiness, we wouldn't immediately throw out our theory that money causes happiness. Instead, the fault might be in any one of our auxiliary assumptions, like our measurement of happiness, or our choice of how much money to give or when to give it. The idea that individual parts of a theory can't be falsified independently is sometimes called holism.

One consequence of holism is that the relationship between data and theory isn't always straightforward. An unexpected observation may not cause us to give up on a main hypothesis in our theory—but it will often cause us to question our auxiliary assumptions instead (e.g., how we operationalize our constructs). Thus, before abandoning our theory of money causing happiness, we might want to try several happiness questionnaires.

The broader idea of holism is supported by historical and sociological studies of how science progresses, especially in the work of Kuhn (1962). Examining historical evidence, Kuhn found that scientific revolutions didn't seem to be caused by the falsification of a theoretical statement via an incontrovertible observation. Instead, Kuhn described scientists as mostly working within **paradigms**: sets of questions, assumptions, methods, phenomena, and explanatory hypotheses.

Paradigms allow for activities Kuhn described as *normal science*—that is, testing questions within the paradigm, explaining new observations, or modifying theory to fit these paradigms. But normal science is punctuated by periods of *crisis* when scientists begin to question their theory and their methods. Crises don't happen just because a single observation is inconsistent with the current theory. Rather, there will often be a holistic transition to a new paradigm, typically because of a striking explanatory or predictive success—often one that's outside the scope of the current working theory entirely.

In sum, the lesson of holism is that we can't just put our theories in direct contact with evidence and think that they will be supported or overturned. Instead, we need to think about the scope of our theory (in terms of the phenomena and measures it is meant to explain), as well as the auxiliary hypotheses—operationalizations—that link it to specific observations.

2.3 Designing experiments to test theory

Source

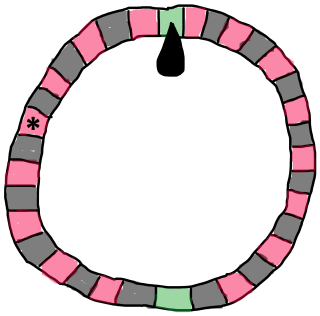


Figure 2.3: A roulette wheel. Betting on red is not that risky, but betting all your chips on a particular value (*) is much riskier.

One way of looking at theories is that they let us make *bets*. If we bet on a spin of the roulette wheel in figure 2.3 that it will show us red as opposed to black, we have almost a 50% chance of winning the bet. Winning such a bet is not impressive. But if we call a particular number, the bet is riskier because we have a much smaller chance of being right. Cases where a theory has many chances to be wrong are called **risky tests** (Meehl 1978).[12]

Much psychology consists of verbal theories. Verbal theories make only qualitative predictions, so it is hard convincingly show them to be wrong (Meehl 1990). In our discussion of money and happiness, we just expected happiness to go up as money increased. We would have accepted *any* increase in happiness (even if very small) as evidence confirming our hypothesis. Predicting that it does is a bit like betting on red with the roulette wheel—it's not surprising or impressive when you win. And in psychology, verbal theories often predict that multiple factors interact with one another. With these theories, it's easy to say that one or the other was "dominant" in a particular situation, meaning you can predict almost any direction of effect.

To test theories, we should design experiments to test conditions where our theories make "risky" predictions. A stronger version of the money-happiness theory might suggest that happiness increases linearly in the logarithm of income (Killingsworth et al. 2023). This specific mathematical form for the relationship—as well as the more specific operationalization of money as income—creates opportunities for making much riskier bets about new experiments. This kind of case is more akin to betting on a specific number on the roulette wheel: when you win this bet, it is quite surprising![13]

Testing theoretical predictions also requires precise experimental measurements. As we start to measure the precision of our experimental estimates in chapter 6, we'll see that the more precise our estimate is, the more values are inconsistent with it. In this sense, a risky test of a theory requires both a specific prediction and a precise measurement. (Imagine spinning the roulette wheel but seeing such a blurry image of the result that you can't really tell where the ball is. Not very useful.)

[12]Even if you're not a falsificationist like Popper, you can still think it's useful to try and falsify theories! Although a single observation is not always enough to overturn a theory, it's still a great research strategy to look for those observations that are most inconsistent with the theory.

[13]Theories are often developed iteratively. It's common to start with a theory that is less precise and, hence, has fewer opportunities for risky tests. By collecting data and testing different alternatives, it's often possible to refine the theory so that it is more specific and allows riskier tests. As we discuss below, formalizing theories using mathematical or computational models is one important route to making more specific predictions and creating riskier tests.

Even when theories make precise predictions, they can still be too flexible to be tested. When a theory has many **free parameters**—numerical values that can be fit to a particular dataset, changing the theories predictions on a case-by-case basis—then it can often predict a wide range of possible results. This kind of flexibility reduces the value of any particular experimental test, because the theorist can always say after the fact that the parameters were wrong but not the theory itself (Roberts and Pashler 2000).

One important way to remove this kind of flexibility is to make predictions in advance, holding all parameters constant. A preregistration is a great way to do this—the experimenter derives predictions and specifies in advance how they will be compared to the results of the experiment. We'll talk much more about preregistration in chapter 11.

We've been focusing mostly on testing a single theory. But the best state of affairs is if a theory can make a very specific prediction that other theories don't make. If competing theories both predict that money increases happiness to the same extent, then data consistent with that predicted relationship don't differentiate between the theories, no matter how specific the prediction might be. The experiment that teaches us the most is going to be the one where a very specific pattern of data is predicted according to one theory and another.[14]

Given all of this discussion, as a researcher trying to come up with a specific research idea, what do you do? Our advice is: *follow the theories*. That is, for the general topic you're interested in—whether it's money and happiness, bilingualism, the nature of concepts, or depression—try to get a good sense of the existing theories. Not all theories will make specific, testable predictions, but hopefully some will! Then ask, what are the "risky bets" that these theories make? Do different theories make different bets about the same effect? If so, that's the effect you want to measure!

The next selections turn from theoretical claims to the choice of a measure.

8.1.1 Measurement scales

Source

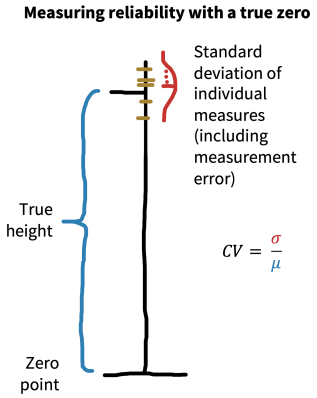


Figure 8.3: Computing the coefficient of variation (CV).

In the physical sciences, it's common to measure the precision of an instrument using its coefficient of variation (Brandmaier et al. 2018):

[14]We can use this idea, which comes from Bayesian statistics, to try to figure out what the *right* experiment is by considering which experimental conditions elicit differences between theories. In fact, the idea of choosing experiments based on the predictions that different theories make has a long history in statistics (Lindley 1956); it's generally called **optimal experiment design** (Myung et al. 2013). The idea is, if you have two or more theories spelled out mathematically or computationally, you can simulate their predictions across a lot of conditions and pick the most informative conditions to run as an actual experiment.

$$\text{CV} = \frac{\sigma_w}{\mu_w}$$

where σ_w is the standard deviation of the measurements within an individual and μ_w is the mean of those measurements (figure 8.3).

Imagine we measure the height of a person five times, resulting in measurements of 171 cm, 172 cm, 171 cm, 173 cm, and 172 cm. These are the combination of the person's true height (we assume they have one!) and some **measurement error**. Now we can use these measurements to compute the coefficient of variation, which is 0.005, suggesting very limited variability relative to the overall quantity being measured. Why can't we just do this same thing with psychological measurements?

Thinking about this question takes us on a detour through the different kinds of measurement scales used in psychological research (Stevens 1946). The height measurements in our example are on what is known as a **ratio** scale: a scale in which numerical measurements are equally spaced and on which there is a true zero point. These scales are common for physical quantities but somewhat less frequent in psychology (with reaction times as a notable exception). More common are **interval** scales, in which there is no true zero point. For example, IQ (and other standardized scores) are intended to capture interval variation on some dimension but zero is meaningless—an IQ of zero does not correspond to any particular interpretation.[15]

Ordinal scales are also often used. These are scales that are ordered but are not necessarily spaced equally. For example, levels of educational achievement ("elementary," "high school," "some college," "college," "graduate school") are ordered, but there is no sense in which "high school" is as far from "elementary" as "graduate school" is from "college." The last type in Stevens's hierarchy is **nominal** scales, in which no ordering is possible either. For example, race is an unordered scale in which multiple categories are present but there is no inherent ordering of these categories. The hierarchy is shown in table 8.1.

Scale Definition Operations Statistics Nominal Unordered list Equality Mode Ordinal Ordered list Greater than or less than Median Interval Numerical Equality of intervals Mean, SD Ratio Numerical & zero Equality of ratios Coefficient of variation Table 8.1: Measurement scale types and their associated operations and statistics (Stevens 1946). Critically, different summary measures work for each scale type. If you have an unordered list like options for a question about race on a survey, you can present the modal (most likely) response. It doesn't even make sense to think about what the median was—there's no ordering! For ordered levels of education, a median is possible but you can't compute a mean. And for interval variables like "number of correct answers on a math test," you can compute a mean and a standard deviation.[16]

Now we're ready to answer our initial question about why we can't quantify reliability using the coefficient of variation. Unless you have a ratio scale with a true zero, you can't compute a coefficient of variation. Think about it for IQ scores: currently, by convention, standardized IQ scores are set to have a mean of 100. If we tested someone multiple times and found the standard deviation of their test scores was four points, then we could estimate the precision of their measurements as "CV" of $4/100 = 0.04$. But since IQ of zero isn't meaningful, we could just set the mean IQ for the

[15] It can actually be shown in a suitably rigorous sense that ratio and interval scales (and another lying in between) are the *only* scales possible for the real numbers (Narens and Luce 1986).

[16] You might be tempted to think that "number of correct answers" is a ratio variable—but is zero really meaningful? Does it truly correspond to "no math knowledge" or is it just a stand-in for "less math knowledge than this test requires"?

population to 200. Our test would be the same, and so the CV would be $4/200 = 0.02$. On that logic, we just doubled the precision of our measurements by rescaling the test! That doesn't make any sense.

8.1.2 Measuring reliability

Source

So then, how do we measure signal and noise when we don't have a true zero? We can still look at the variation between repeated measurement, but rather than comparing that variation between measurements to the mean, we can compare it to some other kind of variation—for example, variation between people. In what follows, we'll discuss reliability on interval scales, but many of the same tools have been developed for ordinal and nominal scales.

Imagine that you are developing an instrument to measure some cognitive ability. We assume that every participant has a true ability, t, just the same way that they have a true height in the example above. Every time we measure this true ability with our instrument, however, it gets messed up by some measurement error. Let's specify that error is normally distributed with a mean of zero—so it doesn't **bias** the measurements, it just adds noise. The result is our observed score, o.[17]

Taking this approach, we could define a relative version of the coefficient of variation. The idea is that the reliability of a measurement is the amount of variance attributable to the true score variance (signal), rather than the observed score variance (which includes noise). If σ_t^2 is the variance of the true scores and σ_o^2 is the variance of the observed scores, then this ratio is:

$$R = \frac{\sigma_t^2}{\sigma_o^2}.$$

When noise is high, then the denominator is going to be big and R will go down to 0; when noise is low, the numerator and the denominator will be almost the same and R will approach 1.

This all sounds great, except for one problem: we can't compute reliability using this formula without knowing the true scores and their variance. But if we did, we wouldn't need to measure anything at all!

There are two main approaches to computing reliability from data. Each of them makes an assumption that lets you circumvent the fundamental issue that we only have access to observed scores and not true scores. Let's think these through in the context of a math test.

Test-retest reliability. Imagine you have two parallel versions of your math test that are the same difficulty. Hence, you think a student's score on either one will reflect the same true score, modulo some noise. In that case, you can use these two sets of observed scores (o_1 and o_2) to compute the reliability of the instrument by simply computing the correlation between them (ρ_{o_1,o_2}). The logic is that, if both variants reflect the same true score, then the shared variance (**covariance** in the sense of chapter 5) between them is just σ_t^2, the true score variance, which is the variable that we wanted but didn't have. Test-retest reliability is thus a very convenient way to measure reliability (figure 8.4).

[17] The approach we use to introduce this set of ideas is called **classical test theory**. There are other—more modern—alternative approaches, but CTT (as it's called) is a good starting point for thinking through the concepts.

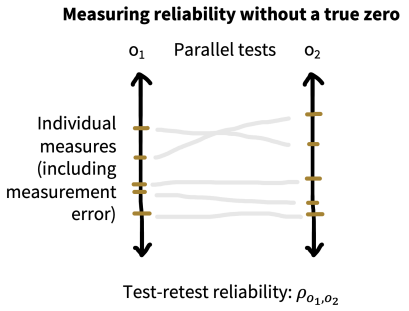


Figure 8.4: Computing test-retest reliability.

Internal reliability. If you don't have two parallel versions of the test, or you can't give the test twice for whatever reason, then you have another option. Assuming you have multiple questions on your math test (which is a good idea!), then you can split the test in pieces and treat the scores from each of these subparts as parallel versions. The simplest way to do this is to split the instrument in half and compute the correlation between participants' scores on the two halves—this quantity is called **split half reliability**.[18]

Another method for computing the internal reliability (**consistency** of a test) is to treat each item as a subinstrument and compute the split-half correlation averaged over all splits. This method yields the statistic **Cronbach's** α ("alpha"). α is widely reported, but also widely misinterpreted (Sijtsma 2009). First, it's actually a lower bound on reliability rather than a good estimate of reliability itself. Second, it's often misinterpreted as evidence that an instrument yields scores that are "internally consistent," which it doesn't; it's not an accurate summary of dimensionality. α is a standard statistic, but it should be used with caution.

One final note: the concept of reliability often gets used to quantify agreement between observers' ratings of the same stimulus (**inter-rater** or **inter-annotator reliability**), say for example when you have two coders rate how aggressive a person seems in a video. The most common measure of inter-annotator agreement is a categorical measure called **Cohen's** κ ("kappa"), for categorical agreement, but you can use **intra-class correlation coefficients** (see the Depth box below) for continuous data as well as many other measures.

8.1.3 Practical advice for computing reliability

Source

If you don't know the reliability of your measures for an experiment, you risk wasting your and your participants' time. Ignorance is not bliss. A higher reliability measure will lead to more precise measurements of a causal effect of interest and, hence, smaller required sample sizes.

Test-retest reliability is generally the most conservative practical measure of reliability. Test-retest estimates include not only measurement error but also participants' state variation across different testing sessions and variance due to differences between versions of your instrument. These real-world quantities are absent from internal reliability estimates, which may make you erroneously think that there is more signal present in your instrument than there is. Even though α is a theoretical lower bound on reliability, in practice, test-retest accuracy often ends up lower than α because it incorporates all these other sources of variation. It's hard work to measure test-retest

[18]The problem is that each half is … half as long as the original instrument. To get around this, there is a correction called the Spearman-Brown correction that can be applied to estimate the expected correlation for the full-length instrument. You also want to make sure that the test doesn't get harder from the beginning to the end. If it does, you may want to use the even-numbered and odd-numbered questions as the two parallel versions.

reliability estimates, in part because you need two different versions of a test (to avoid memory effects). If you plan on using an instrument more than once or twice, though, it will likely be worthwhile!

Finally, if you have multiple measurement items as part of your instrument, make sure you evaluate how they contribute to the reliability of the instrument. Perhaps you have several questions in a survey that you'd like to use to measure the same construct; perhaps multiple experimental vignettes that vary in content or difficulty. Some of these items may not contribute to your instrument's reliability—and some may even detract. At a bare minimum, you should always visualize the distribution of responses across items to scan for **floor and ceiling effects**—when items always yield responses bunched at the bottom or top of the scale, limiting their usefulness—and take a look at whether there are particular items on which items do not relate to the others.

If you are thinking about developing an instrument that you use repeatedly, it may be useful to use more sophisticated psychometric models to estimate the dimensionality of responses on your instrument as well as the properties of the individual items. If your items have binary answers, like test questions, then **item response theory** is a good place to start (Embretson and Reise 2013). If your items are more like ratings on a continuous (interval or ratio) scale, then you may want to look at factor analysis and related methods (Furr 2021).

8.2 Validity

Source (opening paragraphs only)

In chapter 2, we talked about the process of theory building as a process of describing the relationships between constructs. But for the theory to be tested, the constructs must be measured so that you can test the relationships between them! Measurement and measure construction is therefore intimately related to theory construction, and the notion of validity is central.[19]

A valid instrument measures the construct of interest. In figure 8.1, invalidity is pictured as bias—the holes in the target are tightly grouped but in the wrong place.[20] How can you tell if a measure is valid, given that the construct of interest is unobserved? There is no single test of the validity of a measure (Cronbach and Meehl 1955). Rather, the measure is valid if there is evidence that fits into the broader theory as it relates to the specific construct it is supposed to be measuring. For example, it should be strongly related to other measures of the construct, but not as related to measures of different constructs.

How do you establish that a measure fits into the broader theory? Validity of a measure is typically established via an argument that calls on different sources of support (Kane 1992). Here are some of the ways you might support the relationship between a measure and a construct:

- **Face validity**: The measure looks like the construct, such that intuitively it is reasonable that it measures the construct. Face validity is a relatively weak source of evidence for validity, since it relies primarily on pretheoretic intuitions rather than any quantitative assessment. For example, reaction time is typically correlated with intelligence test results (e.g., Jensen and Munro 1979) but

[19]Some authors have treated "validity" as a broader notion that can include, for example, statistical issues (Shadish et al. 2002). The sense of validity that we are interested in here is a bit more specific. We focus on **construct validity**, the relationship between the measure and the construct.

[20]This metaphor is a good rough guide but it doesn't distinguish an instrument that is systematically biased (for example, by estimating scores too low for one group) and one that is invalid (because it measures the wrong construct).

does not appear to be a face-valid measure of intelligence in that simply being fast doesn't accord with our intuition about what it means to be intelligent!

- **Ecological validity**: The measure relates to the context of people's lives. For example, a rating of a child's behavioral self-control in the classroom is a more ecologically valid measure of executive function than a reaction-time task administered in a lab context. Ecological validity arguments can be made on the basis of the experimental task, the stimuli, and the general setting of the experiment (Schmuckler 2001). Researchers differ in how much weight they assign to ecological validity based on their goals and their theoretical orientation.
- **Internal validity**: Usually used negatively. A "challenge to internal validity" is a description of a case where the measure is administered in such a way as to weaken the relationship between measure and construct. For example, if later items on a math test showed lower performance due to test-taker's fatigue rather than lower knowledge of the concepts, the test might have an internal validity issue.[21]
- **Convergent validity**: The classic strategy for showing validity is to show that a measure relates (usually, correlates) with other putative measures of the same construct. When these relationships are measured concurrently, this is sometimes called **concurrent validity**. As we mentioned in chapter 2, self-reports of happiness relate to independent ratings by friends and family, suggesting that both measure the same underlying construct (Sandvik et al. 1993).[22]
- **Predictive validity**. If the measure predicts other later measures of the construct, or related outcomes that might be of broader significance. Predictive validity is often used in lifespan and developmental studies where it is particularly prized for a measure to be able to predict meaningful life outcomes such as educational success in the future. For example, classroom self-control ratings (among other measures) appear strongly predictive of later life health and wealth outcomes (Moffitt et al. 2011).
- **Divergent validity**. If the measure can be shown to be distinct from measure(s) of a different construct, this evidence can help establish that the measure is specifically linked to the target construct. For example, measures of happiness (specifically, life satisfaction) can be distinguished from measures of optimism as well as both positive and negative affect, suggesting that these are distinct constructs (Lucas et al. 1996).

8.2.2 Avoid questionable measurement practices!

Source

Experimentalists sometimes have a tendency to make up ad hoc measures on the fly. When they do so, it's important that they justify their choice in terms of reliability and validity. Table 8.2 gives a set of questions to guide thoughtful reporting of measurement practices.

Question Information to Report What is your construct? Define construct, describe theory and research. What measure did you use to operationalize your construct? Describe measure and justify

[21] Often this concept is described as also being relevant to the validity of a *manipulation*, for example, when the manipulation of the construct is confounded and some other psychological variable is manipulated as well. We discuss internal validity further in chapter 9.

[22] This idea of convergent validity relates to the idea of holism we described in chapter 2. A measure is valid if it relates to other valid measures, which themselves are only valid if the first one is! The measures are valid because the theory works, and the theory works because the measures are valid. This circularity is a difficult but perhaps unavoidable part of constructing psychological theories (see the Depth box above on the history of measurement). We don't ever have an objective starting point for the study of the human mind.

operationalization. Did you select your measure from the literature or create it from scratch? Justify measure selection and review evidence on reliability and validity (or disclose the lack of such evidence). Did you modify your measure during the process? Describe and justify any modifications; note whether they occurred before or after data collection. How did you quantify your measure? Describe decisions underlying the calculation of scores on the measure; note whether these were established before or after data collection and whether they are based on standards from previous literature. Table 8.2: Questions about measurement that every reseacher should answer in their paper. Adapted from Flake and Fried (2020). One big issue to be careful about is that researchers have been known to modify their scales and their scale scoring practices (say, omitting items from a survey or rescaling responses) after data collection. This kind of post hoc alteration of the measurement can sometimes be justified by the data, but it can also look a lot like *p*-hacking! If researchers modify their measurement strategy after seeing their data, this decision needs to be disclosed, and it may undermine their statistical inferences.

8.3 How to select a good measure?

Source

Ideally you want a measure that is reliable and valid. How do you get one? An important first principle is to use a preexisting measure. Perhaps someone else has done the hard work of compiling evidence on reliability and validity, and in that case, you will most likely want to piggyback on that work. Standardized measures are typically broad in their application, and so the tendency can be to discard these because they are not tailored for our studies specifically. But the benefits of a standardized measure are substantial. Not only can you justify the measure using the prior literature but you also have an important index of population variability by comparing absolute scores to other reports.[23]

If you don't use someone else's measure, you'll need to make one up yourself. Most experimenters go down this route at some point, but if you do, remember that you will need to figure out how to estimate its reliability and also how to make an argument for its validity!

We can assign numbers to almost anything people do. We could run an experiment on children's exploratory play and count the number of times they interact with another child (Ross and Lollis 1989), or run an experiment on aggression where we quantify the amount of hot sauce participants serve (Lieberman et al. 1999). Yet, most of the time we choose from a relatively small set of operational variables: asking survey questions, collecting choices and reaction times, and measuring physiological variables like eye movements. Besides following these conventions, how do we choose the right measurement type for a particular experiment?

There's no hard and fast rule about what aspect of behavior to measure, but here we will focus on two dimensions that can help us organize the broad space of possible measure targets.[24] The first of these is the continuum between simple and complex behaviors. The second is the focus on explicit, voluntary behaviors vs implicit or involuntary behaviors.

[23]Comparing absolute measurements is a really important trick for "sanity-checking" your data. If your measurements are very different than the ones in the paper you're following up (for example, if reaction times are much longer or shorter, or if accuracies on a test are much higher or lower), that may be a signal that something has gone wrong.

[24]Some authors differentiate between "self-report" and "observational" measures. This distinction seems simple on its face but actually gets kind of complicated. Is a facial expression a "self-report"? Language is not the only way that people communicate with one another—many actions are intended to be communicative (Shafto et al. 2012).

8.3.1 Simple vs complex behaviors

Figure 8.5 shows a continuum between simple and complex behaviors. The simplest measurable behaviors tend to be button presses, such as:

- pressing a key to advance to the next word in a word-by-word self-paced reading study;
- selecting "yes" or "no" in a lexical decision task; and
- making a forced choice between different alternatives to indicate which has been seen before.

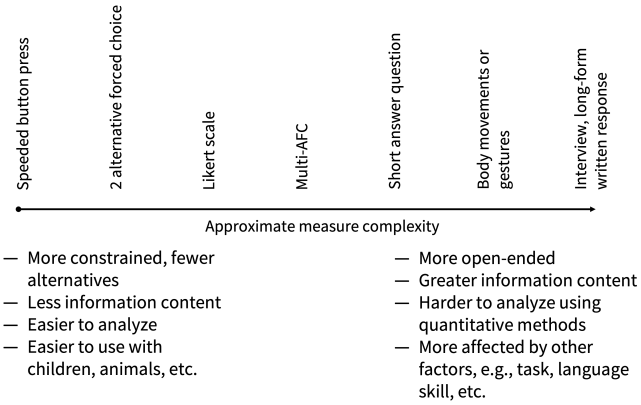


Figure 8.5: Often choosing a measure can be consolidated into a choice along a continuum from simple measures that provide a small amount of information but are quick and easy to repeat and those that provide much richer information but require more time.

These specific measures—and many more like them—are the bread and butter of many cognitive psychology studies. Because they are quick and easy to explain, these tasks can be repeated over many trials. They can also be executed with a wider variety of populations including with young children and sometimes even with nonhuman animals with appropriate adaptation. (A further benefit of these paradigms is that they can yield useful reaction time data, which we discuss further below.)

In contrast, a huge range of complex behaviors have been studied by psychologists, including:

- open-ended verbal interviews;
- written expression, for example, via handwriting or writing style;
- body movements, including gestures, walking, or dance; and
- drawing or artifact building.

There are many reasons to study these kinds of behaviors. First, the behaviors themselves may be examples of tasks of interest (e.g., studies of drawing that seek to understand the origins of artistic expression). Or, the behavior may stand in for other even more complex behaviors of interest, as in studies of typing that use this behavior as a proxy for lexical knowledge (Rumelhart and Norman 1982).

Complex behaviors typically afford a huge variety of different measurement strategies. So any experiment that uses a particular measurement of a complex behavior will typically need to do significant work up front to justify the choice of that measurement strategy—for example, how to quantify dances or gestures or typing errors—and provide some assurance about its reliability. Further, it is often much more difficult to have a participant repeat a complex behavior many times under the same conditions. Imagine asking someone to draw hundreds of sketches as opposed to pressing a key hundreds of times! Thus, the choice of a complex behavior is often a choice to forego a large number of simple trials for a small number of more complex trials.

Complex behaviors can be especially useful to study either at the beginning or the end of a program of research. At the beginning of a set of experiments, they can provide inspiration about the richness of the target behavior and insight into the many factors that influence it. And at the end, they can provide an ecologically valid measure to complement a reliable but more artificial lab-based behavior.

The more complex the behavior, however, the more it will vary across individuals and the more environmental and situational factors will affect it. These can be important parts of the phenomenon, but they can also be nuisances that are difficult to get under experimental control.[25] Simple measures are typically easier to use and, hence, easier to deploy repeatedly in a set of experiments where you iterate your manipulation to test a causal theory.

8.3.2 Implicit vs explicit behaviors

A second important dimension of organization for measures is the difference between implicit and explicit measures. An explicit measure provides a measurement of a behavior that a participant has conscious awareness of—for example, the answer to a question. In contrast, implicit measures provide measurements of psychological processes that participants are unable to report (or occasionally, unwilling to).[26] Implicit measures, especially reaction time, have long been argued to reflect internal psychological processes (Donders 1969 [1868]). They also have been proposed as measures of qualities such as racial bias that participants may have motivation not to disclose (Greenwald et al. 1998). There are also of course a host of physiological measurements available. Some of these measure eye movements, heart rate, or skin conductance, which can be linked to aspects of cognitive process. Others reflect underlying brain activity via the signals associated with MRI (magnetic resonance imaging), MEG (magnetoencephalography), NIRS (near-infrared spectroscopy), and EEG (electroencephalogram) measurements. These methods are outside the scope of this book, though we note that the measurement concerns we discuss here definitely apply (e.g., Zuo et al. 2019).

Many tasks produce both accuracy and reaction time data. Often these trade off with one another in a classic **speed-accuracy trade-off**: the faster participants respond, the less accurate they are. For example, to investigate racial bias in policing, Payne (2001) showed US college students a series of pictures of tools and guns, proceeded by a prime of a White face or a Black face. In a first study, participants were faster to identify weapons when primed by a Black face but had similar accuracies. A second study added a response deadline to speed up judgments: this resulted in equal reaction times across conditions but greater errors in weapon identification after Black faces. These studies likely revealed the same phenomenon—some sort of bias to associate Black faces with weapons—but the design of the task moved participants along a speed-accuracy trade-off, yielding effects on different measures.[27]

---

[25] When they are not designed with care, complex, open-ended behaviors such as verbal interviews can be especially affected by the experimental biases that we will describe in chapter 9, including, for example, **demand characteristics**, in which participants say what they think experimenters would want to hear. Qualitative interview methods can be incredibly powerful as a method in their own right, but they should be deployed with care as measures for an experimental intervention.

[26] Implicit/explicit is likely more of a continuum, but one cut-point is whether the participants' behavior is considered intentional: that is, participants *intend* to press a key to register a decision, but they likely do not intend to react in 300 as opposed to 350 milliseconds due to having seen a prime.

[27] One way of describing the information processing underlying this trade-off is given by drift diffusion models (DDMs), which allow joint analysis of accuracy and reaction time (Voss et al. 2013). Used appropriately, DDMs can provide a way to remove speed-accuracy trade-offs and extract more reliable signals from tasks where accuracy and reaction time are both measured (see Johnson et al. 2017 for an example of a DDM on a weapon-decision task).

Simple, explicit behaviors are often a good starting point. Work using these measures—often the least ecologically valid—can be enriched with implicit measures or measurements of more complex behaviors.

8.4 The temptation to measure lots of things

Source

If one measure is good, shouldn't two be better? Many experimenters add multiple measurements to their experiments, reasoning that more data is better than less. But that's not always true!

Deciding whether to include multiple measures is an aesthetic and practical issue as well as a scientific one. Throughout this book, we have been advocating for a viewpoint in which experiments should be as simple as possible. For us, the best experiment is one that shows that a simple and valid manipulation affects a single measure that is reliable and valid.[28] If you are tempted to include more than one measure, see if we can talk you out of it.[29]

First, make sure that including more measures doesn't compromise each individual measure. This can happen via fatigue or carryover effects. For example, if a brief attitude manipulation is followed by multiple questionnaire measures, it is a good bet that there is likely to be "fade-out" of the effect over time, so it won't have the same effect on the first questionnaire as the last one. Further, even if a manipulation has a long duration effect on participants, survey fatigue may lead to less meaningful responses to later questions (Herzog and Bachman 1981).

Second, consider whether you have a strong prediction for each measure, or whether you're just looking for more ways to see an effect of your manipulation. As discussed in chapter 2, we think of an experiment as a "bet." The more measures you add, the more bets you are making but the less value you are putting on each. In essence, you are "hedging your bets," and so the success of any one bet is less convincing.

Third, if you include multiple measures in your experiment, you need to think about how you will interpret inconsistent results. Imagine you have experimental participants engage in a brief written reflection that is hypothesized to affect a construct (vs a control writing exercise, say listing meals). If you include two measures of the construct of interest and one shows a larger effect, what will you conclude? It may be tempting to assume that the one that shows a larger effect is the "better measure," but the logic is circular—it's only better if the manipulation affected the construct of interest, which is what you were testing in the first place! Including multiple measures because you're uncertain which one is more related to the construct indulges in this circular logic, since the experiment often can't resolve the situation. A much better move in this case is to do a preliminary study of the reliability and validity of the two measures so as to be able to select one as the experiment's primary endpoint.[30]

---

[28] In an entertaining article called "Things I Have Learned (So Far)," Cohen (1990) quips that he leans so far in the direction of large numbers of observations and small numbers of measures that some students think his perfect study has 10,000 participants and *no* measures.

[29] As usual, we want to qualify that we are only talking about randomized experiments here! In observational studies, often the point is to measure the associations between multiple measures, so you typically *have* to include more than one. Additionally, some of the authors of this book have advocated for measuring multiple outcomes in longitudinal observational studies, which could reduce investigator bias, encourage reporting null effects, enable comparison of effect sizes, and improve research efficiency (VanderWeele et al. 2020). We've also done plenty of descriptive studies—these can be very valuable. In a descriptive context, often the goal is to include as many measures as possible so as to have a holistic picture of the phenomenon of interest.

[30] One caveat to this argument is that it can sometimes be useful to examine the effects of a manipulation on different measures because the measures are important. For example, you might be interested in whether an educational intervention increased grades *and* decreased dropout rates. Both outcome measures are important, and so it is useful to include both in your study.

Finally, if you do include multiple measures, selective reporting of significant or hypothesis-aligned measures becomes a real risk. For this reason, preregistration and transparent reporting of all measures becomes even more important.

There are some cases where more measures are better. The more expensive the experiment, the less likely it is to be repeated to gather a new measurement of the effects of the same manipulation. Thus, larger studies present a stronger rationale for including multiple measures. Clinical trials often involve interventions that can have effects on many different measures; imagine a cancer treatment that might affect mortality rates, quality of life, tumor growth rates, and other measures. Further, such trials are extremely expensive and difficult to repeat. Thus, there is a good reason for including more measures in such studies.

The final selections introduce effect estimates and the uncertainty around them.

5.2.1 Estimating the treatment effect

Source

Let's refer to the causal effect we care about as our **treatment effect**.[31] In practice, estimating β (a parameter describing the treatment effect) is going to be a pretty straightforward extension to what we did before.

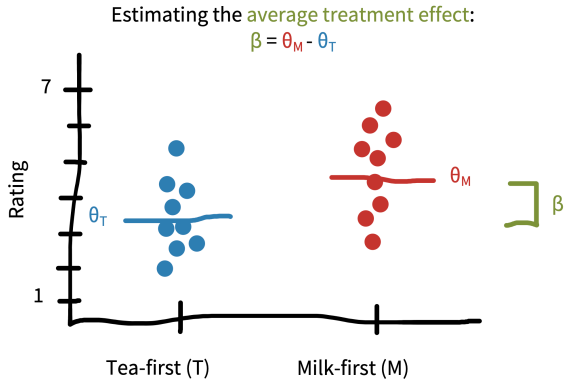


Figure 5.10: Estimating the average treatment effect from the tea-tasting data.

In the maximum likelihood framework, we could posit that ratings in each group (milk-first and tea-first) follow a normal distribution but that these normal distributions might have different means and standard deviations. Extending the notation introduced above, let's term the parameters for the tea-first group θ_T (the mean) and σ (the standard deviation). To estimate the treatment effect, we are positing a model in which the milk-first ratings are normally distributed with mean $\theta_M = \theta_T + \beta$ and with standard deviation σ.[32] This equation says that milk-first ratings have the same distribution as tea-first ratings, except that their average is shifted by β. Setting our model up this way then lets us compute $\hat{\beta}$, our estimate of the treatment effect in our sample.

As in the one-sample case (estimating the mean of just the milk-first group), maximum likelihood estimation would then proceed by finding the value of β that makes the data most likely under the assumed model. As you might expect, this estimate $\hat{\beta}$ turns out to be simply the difference in sample means, $\hat{\theta}_M - \hat{\theta}_T$. This difference is pictured in figure 5.10.

[31] This is the effect of our manipulation—what we sometimes call an "intervention" as well. "Treatment" is a term that comes from medical statistics but is used more broadly in statistics now.

[32] For simplicity, we're assuming that the standard deviations in each tea group are equal.

In the Bayesian framework, we would again specify a prior $p(\beta)$ that encodes our prior beliefs about the size and direction of the treatment effect. If we have no prior beliefs at all, then we could specify a flat prior, $p(\beta) \propto 1$.[33] If we believe the treatment effect is likely to favor milk-first pouring ($\beta > 0$), we could specify that the prior is a normal distribution centered at some positive value (e.g., $\beta = 0.5$); the standard deviation of this prior would encode how certain we are about our prior beliefs. And, if we have no prior beliefs about the direction of the treatment effect, but we think it is unlikely to be very large, we could specify a normal prior centered at 0, which has the effect of "shrinking" the estimates closer to 0.[34]

As in our example above, maximum likelihood estimates and Bayesian estimates are going to be pretty similar if we have a lot of data or weak priors. They will only diverge when we have strong priors or relatively little data. The reason we are setting up these two different frameworks, however, is that they provide very different inferential tools, as we'll see in the next chapter.

6.1.1 Standard errors

Source

To characterize the uncertainty in an estimate, it helps to picture its **sampling distribution**, which is the distribution of the estimate across different, hypothetical, samples. That is, let's imagine that we conducted the tea experiment not just once but dozens, hundreds, or even thousands of times. This idea is often called **repeated sampling** as a shorthand. For each hypothetical sample, we use similar recruitment methods to recruit a new sample of participants, and we compute $\hat{\beta}$ for that sample. Would we get exactly the same answer each time? No, simply because the samples will have some random variability (noise). If we plotted these estimates, $\hat{\beta}$, we would get the sampling distribution in figure 6.2.

> In this chapter and the subsequent statistics and visualization chapters of the book, we'll try to facilitate understanding and illustrate how to use these concepts in practice by giving the R code we use in constructing our examples in these **code** boxes. We'll assume that you have some knowledge of base R and the Tidyverse—to get started with these, go ahead and take a look at appendix D if you haven't already. Although our figures are often drawn by hand, even the hand-drawn ones are based on actual simulation results!
>
> Since we're going to be working with lots of data from the tea tasting example, we wrote a function called `make_tea_data()` that creates a `tibble` with some (made-up) data from our modern tea-tasting experiment. You can find the function on GitHub (https://github.com/langcog/experimentology/blob/main/helper/tea_helper.qmd) if you want to follow along.

```
tea_data <- make_tea_data(n_total = 18)
```

[33] This equation says that the probability of any value of β is "proportional to" 1, meaning that it's constant ("flat") regardless of what value β takes.

[34] The measures of variability that we discuss here account for statistical uncertainty reflecting the fact that we have only a finite sample size. If the sample size were infinite, there would be no uncertainty of this kind. Statistical uncertainty is only one kind of uncertainty, though. A more holistic view of the overall credibility of an estimate should also account for other things outside of the model, like study design issues and bias.

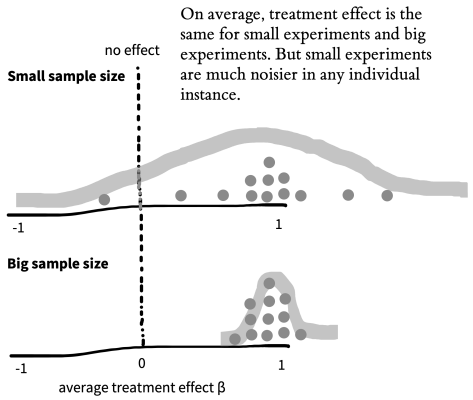


Figure 6.3: Comparing sampling distributions for the treatment effect by sample size.

Now imagine we also did thousands of repetitions of the experiment with $n = 200$ per group instead of $n = 9$ per group. Figure 6.3 shows what the sampling distribution might look like in that case. Notice how much narrower the sampling distribution becomes when we increase the sample size, showing our decreased uncertainty. More formally, the standard deviation of the sampling distribution itself, called the **standard error**, decreases as the sample size increases.

The sampling distribution is not the same thing as the distribution of tea ratings in a single sample. Instead, it's a distribution of *estimates across samples of a given size*. In essence, it tells us what the mean of a new experiment might be, if we ran it with a particular sample size.

To do simulations where we repeat the tea-tasting experiment over and over again, we're using a special tidyverse function from the `purrr` library: `map()`. `map()` is an extremely powerful function that allows us to run another function (in this case, the `make_tea_data()` function that we introduced last chapter) many times with different inputs. Here we create a tibble made up of a set of 1,000 runs of the `make_tea_data()` function.

```
samps <- tibble(sim = 1:1000) |>
  mutate(data = map(sim, \(i) make_tea_data(n_total = 18))) |>
  unnest(cols = data)
```

Next, we just use the `group_by()` and `summarize()` workflow from appendix D to get the estimated treatment effect for each of these simulations.

```
tea_summary <- samps |>
  group_by(sim, condition) |>
  summarize(mean_rating = mean(rating)) |>
  group_by(sim) |>
  summarize(delta = mean_rating[condition == "milk first"] -
              mean_rating[condition == "tea first"])
```

This tibble gives us what we would need to plot the sampling distributions above in figure 6.2 and figure 6.3.

6.5.1 Confidence intervals

Source

One alternative to binary hypothesis testing is to ask about the precision of our estimates—how similar an estimate from a particular sample is to the population parameter of interest. For example, how close is our tea-tasting effect estimate to the true effect in the population? We don't know what the true effect is, but knowledge of sampling distributions lets us make some guesses about how precise our estimate is.

The **confidence interval** is a convenient frequentist way to summarize the variability of the sampling distribution—and hence how precise our point estimate is. The confidence interval represents the range of possible values for the parameter of interest that are plausible given the data. More formally, a 95% confidence interval for some estimate (call it $\hat{\beta}$, as in our example) is defined as a range of possible values for β such that, if we did repeated sampling, 95% of the intervals generated by those samples would contain the true parameter, β.

Confidence intervals (CIs) are constructed by estimating the middle 95% of the sampling distribution of $\hat{\beta}$. Because of our hero, the central limit theorem, we can treat the sampling distribution as normal for reasonably large samples. Given this, it's common to construct a 95% confidence interval $\hat{\beta} \pm 1.96\ \widehat{\text{SE}}$.[35] If we were to conduct the experiment 100 times and calculate a confidence interval each time, we should expect 95 of the intervals to contain the true β, whereas we would expect the remaining 5 to not contain β.[36]

Confidence intervals are like betting on the inferences drawn from your sample. The sample you drew is like one pull of a slot machine that will pay off (i.e., have the confidence interval contain the true parameter) 95% of the time. Put more concisely: 95% of 95% confidence intervals contain the true value of the population parameter.

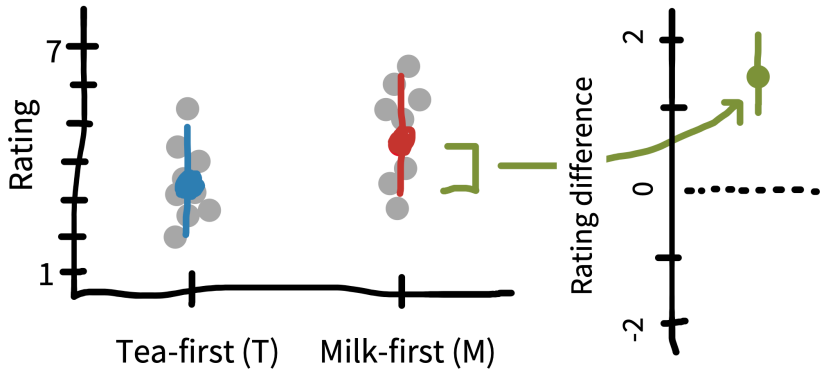


Figure 6.11: Confidence intervals on each of the two condition estimates, as well as on the difference between conditions.

For visualization, we can show the confidence intervals on individual estimates (left side of figure 6.11). These tell us about the precision of our estimates of each quantity relative to the population estimate. But we've been talking primarily about the CI on the treatment effect $\hat{\beta}$ (right side of figure 6.11). This CI allows us to make an inference about whether or not it overlaps with zero—which is actually equivalent in this case to whether or not the t-test is statistically significant.

[35] This type of CI is called a "Wald" confidence interval.

[36] In case you don't have enough tea to do the experiment 100 times to confirm this, you can do it virtually using this nice simulation tool: https://istats.shinyapps.io/ExploreCoverage.

Computing confidence intervals analytically is pretty easy. Here we first compute the standard error for the difference between conditions. The only tricky bit here is that we need to compute a pooled standard deviation.

```
tea_ratings <- filter(tea_data, condition == "tea first")$rating
milk_ratings <- filter(tea_data, condition == "milk first")$rating

n_tea <- length(tea_ratings)
n_milk <- length(milk_ratings)
sd_tea <- sd(tea_ratings)
sd_milk <- sd(milk_ratings)

tea_sd_pooled <- sqrt(((n_tea - 1) * sd_tea ^ 2 + (n_milk - 1) * sd_milk ^ 2) /
                        (n_tea + n_milk - 2))

tea_se <- tea_sd_pooled * sqrt((1 / n_tea) + (1 / n_milk))
```

Once we have the standard error, we can get the estimated difference between conditions and compute the confidence intervals by multiplying the standard error by 1.96.

```
delta_hat <- mean(milk_ratings) - mean(tea_ratings)
tea_ci_lower <- delta_hat - tea_se * qnorm(0.975)
tea_ci_upper <- delta_hat + tea_se * qnorm(0.975)
```

6.5.2 Confidence in confidence intervals?

Source

Confidence intervals are often misinterpreted by students and researchers alike (Hoekstra et al. 2014). Imagine a researcher conducts an experiment and reports that "the 95% confidence interval for the mean ranges from 0.1 to 0.4." All of the statements in table 6.4, though tempting to make about this situation, are *technically false*.

Misconception 1 "The probability that the true mean is greater than 0 is at least 95%." 2 "The probability that the true mean equals 0 is smaller than 5%." 3 "The 'null hypothesis' that the true mean equals 0 is likely to be incorrect." 4 "There is a 95% probability that the true mean lies between 0.1 and 0.4." 5 "We can be 95% confident that the true mean lies between 0.1 and 0.4." 6 "If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4." Table 6.4: Confidence interval misconceptions for a confidence interval $[0.1, 0.4]$. Adapted from Hoekstra et al. (2014). The problem with all of these statements is that, in the frequentist framework, there is only one true value of the population parameter, and the variability captured in confidence intervals is about the *samples*, not the parameter itself.[37] For this reason, we can't make any statements about the probability of the value of the parameter or of our confidence in specific numbers. To reiterate, what we *can* say is: if we were to repeat the procedure of conducting the experiment and calculating a confidence interval many times, in the long run, 95% of those confidence intervals would contain the true parameter.

The Bayesian analog to a confidence interval is a **credible interval**. Recall that for Bayesians, parameters themselves are considered probabilistic (i.e., subject to random variation), not fixed. A

[37] In contrast, Bayesians think of parameters themselves as variable rather than fixed.

95% credible interval for an estimate, $\hat{\beta}$, represents a range of possible values for β such that there is a 95% probability that β falls inside the interval. Because we are now wearing our Bayesian hats, we are "allowed" to talk about β as if it were probabilistic rather than fixed. In practice, credible intervals are constructed by finding the posterior distribution of β, as in chapter 5, and then taking the middle 95%, for example.

Credible intervals are nice because they don't give rise to many of the inference fallacies surrounding confidence intervals. They actually *do* represent our beliefs about where β is likely to be, for example. Despite the technical differences between credible intervals and confidence intervals, in practice—with larger sample sizes and weaker priors—they turn out to be quite similar to each other in many cases.[38]

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), *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.

[38] They can diverge sharply in cases with less data or stronger priors (Morey et al. 2016), but in our experience this is relatively rare.